Evolution of Microbial Metabolism

Dissertation

zur

Erlangung der naturwissenschaftlichen Doktorwürde

(Dr. sc. nat.)

vorgelegt der

Mathematisch-naturwissenschaftlichen Fakultät

der

Universität Zürich

von

Aditya Barve

aus

Indien

Promotionskomitee:

Prof. Dr. Andreas Wagner (Vorsitz)

Prof. Dr. Olivier Martin

Prof. Dr. Rolf Kümmerli

Zürich, 2014

All that happens to us, including our humiliations, our misfortunes, our embarrassments, all is given to us as raw material, as clay, so that we may shape our art.

-- Jorge Luis Borges

In science, self-satisfaction is death. Personal self-satisfaction is the death of the scientist. Collective self-satisfaction is the death of the research. It is restlessness, anxiety, dissatisfaction, and agony of mind that nourish science.

-- Jacques-Lucien Monod

Abstract

Microbes include some of the most ancient and ubiquitous organisms alive today and have a large impact on human society. They are involved in major geophysical cycles, used in biotechnology, and are one of the main causes of disease in humans. In my thesis I have focused on the evolution of microbial metabolism. I generated and analyzed random metabolisms that are viable on a given carbon source. These metabolisms can synthesize all biomass precursors necessary for growth and survival, but otherwise contain a random set of reactions. One way to study the evolution of metabolism is to study essential reactions, that is, metabolic reactions that are necessary for the survival and growth of an organism. I found that metabolisms viable on different carbon sources contain a core of 125 absolutely superessential reactions, that is, each of these reactions is required for the survival of the organism independently of the carbon source the organism is viable on. I then assigned a superessentiality index to each reaction, which denotes the number of different metabolisms in which that reaction was essential. I showed that the superessentiality index correlates with the number of sequenced prokaryotic genomes that encode an enzyme for that reaction, indicating that my approach can reveal biologically relevant information. In a next analysis, I focused on non-adaptive metabolic traits that may lead to evolutionary innovations, such as the ability of an organism to grow on a new carbon source. Specifically, I found that random metabolisms required to grow on a given carbon source are typically also viable on multiple other carbon sources that were not targets of selection. My observations suggest that non-adaptive traits like these may be important for evolutionary innovations. In a last analysis, I focused on a network of metabolic genotypes that are viable on the same carbon source. Two viable metabolisms on such a genotype network are connected if they can be converted into one another through a sequence of single reaction changes that preserve viability. A genotype network contains genotypes that can vary in properties such as fitness and mutational robustness. Connectedness of a genotype network enables evolution to fine-tune these properties. The study shows that all but the simplest metabolisms, which are metabolisms with few reactions, are connected. This means that metabolism is highly evolvable and that its properties can be fine-tuned through small alterations in its reaction content.

Zusammenfassung

Mikroorganismen beinhalten einige der ältesten und allgegenwärtigsten heute noch lebender Organismen und haben einen grossen Einfluss auf die menschliche Gesellschaft. Sie sind an grossen geophysikalischen Zyklen beteiligt, werden in der Biotechnologie eingesetzt und sind eine der Hauptursachen von Krankheiten beim Menschen. In meiner Doktorarbeit konzentrierte ich mich auf die Evolution der mikrobiellen Stoffwechsel. Ich erstellte und analysierte zufällige Stoffwechselprozesse, welche lebensnotwendig für bestimmte Kohlenstoffquellen sind. Diese Stoffwechselprozesse sind in der Lage alle Biomassenvorstufen zu synthetisieren, die für das Wachstum und das Überleben notwendig sind, ansonsten aber eine regellose Reihe von Reaktionen umfassen. Eine Möglichkeit die Evolution von Mikroorganismen zu erforschen, ist die Untersuchung von lebenswichtigen Reaktionen oder genauer die Untersuchung jener Stoffwechselreaktionen, die für das Überleben und das Wachstum eines Organismus unverzichtbar sind. Ich fand heraus, dass die Stoffwechselprozesse, welche lebensnotwendig für bestimmte Kohlenstoffquellen sind, aus einem Kern von 125 absolut unverzichtbaren Reaktionen bestehen. Das bedeutet, dass jede dieser Reaktionen für das Überleben des Organismus erforderlich ist, unabhängig von der Kohlenstoffquelle welche der Organismus zum Leben braucht. Danach wies ich jeder Reaktion einen Unverzichtbarkeits-Index zu, welcher die Anzahl von verschiedenen Stoffwechselprozessen kennzeichnet bei denen diese Reaktion unverzichtbar war. Ich zeigte, dass der Unverzichtbarkeits-Index mit der Anzahl sequenzierten prokaryotischen Genomen korreliert, welche ein Enzym für die Reaktion kodieren. Dies weist darauf hin, dass mein Vorgehen biologisch relevante Informationen enthüllen kann. In einer weiteren Analyse konzentrierte ich mich auf nicht adaptive Stoffwechsel-Eigenschaften, welche zu evolutionären Innovationen führen können, wie die Fähigkeit eines Organismus auf einer neuen Kohlenstoffquelle zu wachsen. Genauer gesagt fand ich heraus, dass zufällige Stoffwechselprozesse, welche für das Wachstum auf bestimmten Kohlenstoffquellen benötigt werden, auch lebensnotwendig auf mehreren anderen Kohlenstoffquellen sind, die nicht Ziel der Selektion waren. Meine Beobachtungen lassen vermuten, dass diese nicht adaptiven

Abstract

Merkmale wichtig für evolutionäre Innovationen sein könnten. In meiner letzten Analyse konzentrierte ich mich auf ein Netzwerk von Stoffwechsel-Genotypen, die lebensnotwendig für die gleichen Kohlenstoffquellen sind. In einem solchen Genotypen Netzwerk sind zwei lebensnotwendige Stoffwechselprozesse miteinander verbunden, wenn sie durch eine Reihe von Veränderungen einzelner Reaktionen, die das Überleben erhalten, ineinander umgewandelt werden können. Genotypen in einem Genotypen Netzwerk können in Eigenschaften wie Fitness und Mutations-Robustheit variieren. Der Verbindungsgrad eines Genotypen Netzwerkes ermöglicht der Evolution die Feinabstimmungen dieser Eigenschaften. Diese Studie zeigt, dass alle Stoffwechselprozesse, ausser den einfachsten welche nur aus wenigen Reaktionen bestehen, miteinander verbunden sind. Dies bedeutet, dass der Stoffwechsel sehr evolvierbar ist und dass seine Eigenschaften durch kleine Änderungen der Reaktionen angepasst werden kann.

Acknowledgements

I dedicate my thesis to my wife Shalmali, my brother Ameet and my parents Mugdha and Anand. I am forever grateful to Shalmali and Ameet, who convinced me to leave my job and think of taking up a new challenge. Thank you Shalmali, you make everything just perfect for me.

I thank my supervisor Andreas for giving me the opportunity to work in his lab and teaching me much of what I know. I am grateful for the scientific freedom he gave me, and for improving my manuscript drafts. I feel special gratitude towards my friend, and colleague João Rodrigues, for simply being a great friend and playing the role of a bouncing board for my crazy ideas. Thank you Niv, for you taught me to never let go! The office has been different since you two moved out.

A special thanks goes to Olivier Martin, Homayoun Bagheri, Rolf Kümmerli, Areejit Samal and Karthik Raman whose help was instrumental in accomplishing my goals. Rzgar Hosseini, who made it possible to look at “all” of the metabolisms. Thank you Tugce Bilgin, my officemate, for tolerating my whims and loud exclamations of frustration and joy. Thank you Heidi for the German translation. Thanks to the entire Wagner lab for the inspiring scientific discussions, especially Eric, Sinisa, Joshua, Jose, Evandro, Kathleen, Daniel and Sorcha. The coffee was so much better with you guys across the table. Thank you Annette and Christian for all the administrative help.

Thank you Akshay and Anuja, Arnab and Nitika, Benesh and Remmia and Utsav for all the fun we had. A “heavy” shout out to my band mates, Alexandra, Claudia, Samuel, Abhishek, and Rahul. Life has been better with the music. Thank you!

Contents

Abstract ..... **iii**

Zusammenfassung ..... **iv**

Acknowledgements ..... **vi**

Contents ..... **vii**

1. Introduction **1**

1.1 Microbial metabolism **1**

1.2 Evolution of metabolism **3**

1.2.1 Horizontal gene transfer 4

1.3 Genotype-Phenotype maps **6**

1.3.1 Genotype networks 6

1.3.2 Robustness 8

1.3.3 Innovation 9

1.3.4 Latent phenotypes and exaptation 11

1.3.5 Historical Contingency 12

1.4 Computational modeling and simulation of metabolism **13**

1.4.1 Kinetic modeling 14

1.4.2 Flux balance analysis 20

1.4.3 Genome-scale metabolic reconstructions 21

1.4.4 Markov chain Monte Carlo (MCMC) sampling 22

1.5 Applications of genome scale models **23**

1.5.1 Evolutionary applications 23

1.5.2 Biotechnology and Medicine 26

1.6 Thesis outline **29**

1.7 References **31**

2. Superessential reactions in metabolic networks **38**

2.1 Abstract **39**

2.3 Results **43**

2.4 Discussion **58**

2.5 Methods **64**

2.6 Supplementary results 73
2.7 References 90

3. A latent capacity for evolutionary innovation through exaptation in complex metabolic systems 97
3.1 Abstract 98
3.2 Introduction 99
3.3 Results 100
3.4 Discussion 107
3.5 Methods 109
3.6 Supplementary results 117
3.7 References 143

4. Historical contingency does not strongly constrain the gradual evolution of metabolic properties in central carbon and genome-scale metabolisms 147
4.1 Abstract 148
4.2 Introduction 149
4.3 Results 152
4.4 Discussion 173
4.5 Methods 177
4.6 Supplementary results 185
4.7 References 203

Curriculum Vitae 209

1. Introduction

Microorganisms are the most ancient and the most diverse life forms on our planet. They are known to colonize highly different environments, from glaciers to hot vents deep in the earth, oceans and bodies of other organisms. They are important participants of global ecology, major geophysical processes[1,2], and responsible for diseases. They are also used in food and biotechnology. The study of microorganisms involves their characterization, isolation and cultivation in the laboratory. Unfortunately, most microbial species are not cultivable in the laboratory, and less than one percent of all living microbes have been cultured so far[3]. There is a lot to learn from the microorganisms that have been characterized, especially from their metabolism, which is the primary focus of the analyses described here. Although the studies described here are primarily focused on the metabolic evolution of free-living prokaryotes, I use relevant examples from different biological systems and organisms.

1.1 Microbial metabolism

The genome of a microorganism encodes enzymes that catalyze thousands of biochemical reactions in the cell at any given time. These reactions allow a cell to convert nutrients into the building blocks of life, such as lipids (required to build the cell wall), amino acids (used to build proteins), nucleotides (required for the synthesis of DNA and RNA), and various other cofactors required by a cell to catalyze biochemical reactions[4]. These molecules are called biomass precursors, and the biochemical processes that enable their synthesis are collectively referred to as metabolism.

Metabolism can be roughly broken down into three sub-processes (1) nutrient transport and breakdown, (2) central or core metabolism and (3) biosynthesis of biomass molecules. The first category, that of nutrient transport and breakdown involves the different ways in which cells import and break down nutrients such as carbon, nitrogen or sulfur. For example, in the bacterium *Escherichia coli* (*E. coli*), the carbon source glucose, is converted into glucose-6-phosphate which can enter the

glycolysis pathway[5]. D-xylose, another carbon source, is converted into D-xylulose-D-phosphate, which enters the pentose phosphate pathway[5]. Different nutrients are converted into molecules that enter central or core metabolism.

The main function of central or core metabolism is the conversion of the products of nutrient uptake into several metabolic precursor molecules. These include 12 especially important metabolites, such as glucose-6-phosphate, phosphoenolpyruvate, and pyruvate[6,7]. In addition, core metabolism also generates sources of cellular energy, such as adenosine triphosphate (ATP) and sources of reductive power, such as nicotinamides (NADH and NADPH). In addition to the above mentioned precursors pyruvate and acetyl-CoA, ribose-5-phosphate, phosphoenolpyruvate, oxaloacetate, glutamate, glutamine, glucose-6-phosphate, glyceraldehyde-3-phosphate, 3-phospho-D-glycerate, erythrose-4-phosphate and fructose-6-phosphate also function as precursors. These precursors participate in several different biochemical pathways (sets of catalytic reactions) that are linked in a complex manner. These interconnections allow cells to efficiently manage the flow of matter while synthesizing needed precursors[7]. Important pathways that participate in core metabolism include Embden-Meyerhof (glycolysis) pathway, the pentose phosphate pathway, and the tricarboxylic acid (TCA) or the Krebs cycle.

Together, the above-mentioned 12 precursors and sources of energy and reductive power allow metabolism to synthesize biomass molecules such as lipids, nucleotides and amino acids. A free-living microorganism such as *E. coli* requires the synthesis of more than 60 such molecules to sustain life and grow [4,6,8]. The synthesis of any of these 60 molecules proceeds through the 12 precursors. For instance, pyruvate, a metabolite that participates in glycolysis is essential for the synthesis of amino acids such alanine, isoleucine, leucine, and lysine[4,5]. Acetyl-Coenzyme A, which links glycolysis and the TCA cycle, is essential for the synthesis of lipids[4,5]. Cofactors such as ATP and NADH, NADPH are required to facilitate such catalytic processes.

Together, these three categories of metabolic processes form a “bow-tie” architecture[9,10]. That is, the numerous pathways of nutrient conversion “fan in” to central core metabolism, which operates as the core engine of the cell, subsequently

fanning out towards synthesis of many individual biomass molecules. This architecture of metabolism has been shown to be a general picture of metabolism by computational analysis of 65 different microorganisms[11].

While the metabolism of an organism may contain thousands of metabolic reactions, some of these reactions may be non-essential for the synthesis of biomass precursors, while deleting or knocking out other reactions abolishes a cell's ability to synthesize biomass precursors. Such reactions are called essential reactions. In Chapter 2, I analyze the essentiality of reactions across many different metabolisms and identify sets of reactions that are never or rarely essential, frequently essential and always essential.

1.2 Evolution of metabolism

The fundamental unit of evolutionary change in the metabolism of an organism is the enzyme-coding gene. Mutations in such a gene can change the activity of the encoded enzyme by changing its catalytic efficiency. Modest change can occur through mutations in the protein-coding region that reduce the activity of an enzyme without abolishing it completely or increase enzyme activity. Although point mutations affect single nucleotides, they can have large effects on the survival and growth of the cell. For instance, they can increase a cell's fitness in an environment [12,13], allow growth on a new carbon source [14], or help a cell respond to the loss of an important enzyme [15].

Here, I am concerned with changes that affect the reaction complement of an organism and occur over longer time scales. These changes involve the loss and gain of genetic material. For instance, the function of an enzyme can be completely lost through gene deletions. Conversely, organisms can acquire new genetic material through a process called horizontal gene transfer. Some newly acquired genes may code for enzymes with new functions, allowing a cell access to new metabolic traits. These two mechanisms can change the reaction complement of an organism. For example, an average of 64 genes may have been transferred into the *E. coli* genome every million years [16,17]. Different *E. coli* strains differ in about 10 percent of their

genome [18] and on average in 36 percent of their metabolic reaction content [19]. About 21 percent of the genes in proteobacterial genomes originate from recent transfer events [20]. Another potential source of metabolic genes with new functions is gene duplication. However, most recent changes to bacterial metabolism are known to occur through horizontal gene transfer [21]. In sum, the loss and gain of genetic material is an important driver of metabolic change.

1.2.1 Horizontal gene transfer

Horizontal gene transfer occurs when organisms incorporate genetic material from other organisms in the same population and even among different species. Examples of horizontal gene transfer and the new traits it confers are common in bacterial evolution. These include the acquisition of enzyme-coding genes responsible for the degradation of xenobiotic molecules [22,23], dietary polysaccharides [24], and antibiotic resistance (reviewed in [25–27]). The process of horizontal gene transfer requires DNA from the potential donor cell and its uptake by the recipient cell through various mechanisms. The newly acquired DNA must be stably expressed in a manner that benefits the recipient organism. The actual transfer of DNA occurs through three mechanisms, namely natural transformation, conjugation and transduction.

Natural transformation occurs when a bacterial cell takes up naked DNA from the environment. Both linear DNA and plasmids can be subject to uptake. Certain organisms such as *Neisseria gonorrhoeae* and *Haemophilus influenzae* are naturally competent, that is, they have the ability to uptake naked DNA at all times [28]. Most other microorganisms acquire competence in a time-constrained manner dependent on the environment. Approximately 1 percent of the currently known bacteria are thought to be naturally transformable [29]. Among them are human pathogens such as *Helicobacter*, *Neisseria*, *Pseudomonas*, *Staphylococcus* and *Streptococcus* [30]. The ability to take up naked DNA from the environment means that very distantly related organisms can transfer genetic material.

In contrast to natural transformation, the mechanism of conjugation requires that the donor and recipient cells are physically linked through cell-to-cell junctions (a

"bridge") and a pore through which DNA can pass. Conjugative transfer systems are encoded by plasmids and transposons and carry within them the proteins necessary for their excision from the donor, formation of the conjugative bridge, and subsequent transfer to the recipient. Gene transfer through conjugation may be a frequent mechanism in proteobacteria [20]. While conjugative transfer is more frequent in related species because of bacterial immunity systems [31], it need not be limited by relatedness [32,33].

The third mechanism of gene transfer is transduction. It involves replicating bacteriophages that replicate within a donor bacterium and may package bacterial DNA along with viral genes. Such bacteriophages can then infect other bacterial hosts leading to the transfer of DNA between species. This mechanism, like transformation, does not require the donor and the recipient species to be present in the same environment [34]. However, transduction can be limited due to host specificity of a virus.

The above-mentioned mechanisms enable the transfer of DNA between highly distant organisms. However, there are several barriers to horizontal gene transfer. Firstly, the physical distance between the donor and the recipient microorganism influences the physical transfer of DNA. This is especially true for conjugative transfer systems, as the mechanism requires a physical bond between two cells. Indeed, 74 percent of detected horizontal gene transfers (independent of the mechanism of transfer) occurred between donors and recipients sharing the same habitat [20]. Secondly, gene transfer frequency is positively correlated with genome sequence similarity between the recipient and donor. Moreover, the GC content of the donor and recipient species may need to be similar for gene transfer to become successful [20]. Lastly, to allow its expression, a gene has to be inserted near a promoter [35] or inserted along with a promoter that the recipient cell recognizes. The similarity of the GC content of such a promoter to that of the recipient genome may be essential for the expression of the newly acquired gene [36]. In addition, even after a gene has integrated with the recipient's genome, it needs to contribute to survival or an increase in fitness in at least one or more environments that the recipient experiences. Genes that do not benefit the organism are likely to accumulate mutations leading to loss-of-function.

1.3 Genotype-Phenotype maps

In this section, I introduce the concept of a genotype-phenotype map, which helps to understand the evolution of a variety of biological systems. A genotype can be defined as the genetic material (DNA or RNA) of organisms, while a phenotype is an observable property of an organism. Genotype-phenotype maps can be conceptualized at different levels of organization. For example, the genotype of an RNA molecule is its nucleotide sequence, while its secondary structure can be viewed as its phenotype. Higher levels of organization, such as gene regulatory circuits are also amenable to such treatment, wherein genetically encoded interactions between genes can specify a genotype, while a stable gene expression pattern may constitute its phenotype. For metabolism, the genotype can be viewed as an organism's reaction content, and its ability to synthesize biomass precursors in specific environments can be defined as its phenotype. Because evolution proceeds through genotypic change, one must study the collection of all genotypes and their associated phenotypes to understand evolution of biological systems in a systematic manner.

1.3.1 Genotype networks

Different biological systems that have been studied using genotype-phenotype maps show certain similarities. First, genotype spaces can be vast. For example, the genotype space of RNA molecules of length L corresponds to 4^L different nucleotide sequences. In comparison, a phenotype space is typically much smaller, because many genotypes show the same phenotype. For example, computational studies of RNA secondary structure show that the size of phenotype space scales only to 1.8^L, meaning that there exist more than 2^L more sequences than structures [37]. For sequences of 100 amino acids, the number of genotypes is 20^{100} ($\approx 10^{130}$). An important phenotype of an amino acid sequence is the fold that it can form, which is a specific arrangement of α-helices and β-sheets in three-dimensions. In comparison to the number of genotypes, the estimated number of protein folds is 10^4 [38]. Gene regulatory circuits with a given number of N genes can have N^2 pairwise regulatory interactions among genes. If only two types of interactions (activation and repression)

are allowed, the genotype space contains 2^{N^2} genotypes. Conversely, the number of gene expression phenotypes is only 2^N if genes are allowed to switch between an on and off state. For metabolism, a genotype is defined by the presence or absence of reactions. An important metabolic phenotype is the set of carbon sources that allow synthesis of biomass precursors. For more than 5000 metabolic reactions that are currently known to us [39,40], the genotype space consists of 2^{5000} metabolic genotypes, while the number of carbon sources that metabolisms can utilize is much smaller [4,8,41–44].

Second, individual genotypes typically have many neighbors with the same phenotype. For example, two RNA molecules are neighbors if they differ by the smallest possible genetic change, such as a single nucleotide. Two such RNA molecules are then said to be "connected". The fact that many genotypes show the same phenotype and that any genotype can have a many neighbors leads us to an important organizing principle of genotype-phenotype maps - one can reach any genotype from another genotype in the connected set through small genotypic changes without disrupting the phenotype. Such a connected set of genotypes with the same phenotype is referred to as a genotype network [45] or a neutral network [37]. Genotype networks of biological molecules can be vast. For example, approximately 10^{57} amino acid sequences (length $L = 96$) can fold into a structure characteristic of the bacteriophage λ transcriptional repressor [46].

Third, genotype networks (the number of genotypes showing a specific phenotype) typically occupy a small fraction of genotype space. Even though genotype networks can be vast in numbers as discussed above, they are still tiny in comparison to the number of all possible genotypes. The genotype space of RNA molecules of length $L = 50$ comprises 4^{50} or 10^{30} molecules, while the number of genotypes forming a particular structure is 10^{17} [47]. This means that RNA molecules forming this structure constitute a tiny fraction 10^{-13} of the genotype space. For protein sequences folding into the λ repressor, this fraction shrinks further to 10^{-63}. In other words, genotypes forming a given phenotype may be very rare in genotype space, meaning that the probability of randomly choosing a genotype with a specific phenotype from genotype space is extremely small.

1.3.2 Robustness

The robustness of a biological system refers to its ability to maintain its phenotype in the face of perturbations. Perturbations can be genetic, such as mutations, and robustness to such perturbations is called mutational robustness. A non-genetic perturbation, such as a changing environment can also affect an organism's phenotype. Such environmental changes may be more frequent than mutations, but mutational change can have a more permanent effect on a system.

Consider an RNA molecule with a given secondary structure. Mutating any one nucleotide position to any of the remaining three nucleotides would result in three neighboring genotypes. If these neighbors have the same secondary structure, one could say that this genotype is robust to mutations at that nucleotide position. If all of a genotype's neighbors have the same phenotype, the genotype would be maximally robust. Conversely, if none of the neighbors showed the same phenotype, the genotype would be fragile. One can thus quantify mutational robustness of a genotype as the fraction of its neighbors with the same phenotype. In a genotype network, the mutational robustness of a genotype is simply the number of its neighbors. The genotype with the highest number of neighbors is the most robust and vice versa. How robust are biological molecules? Sanjuán *et. al.* studied the folding of RNA molecules belonging to 29 different small plant pathogens called viroids [48]. Viroids are single-stranded RNAs and do not code for any protein molecules. They found that the mutational robustness among these 29 RNA genotypes ranges between 0.17 and 0.26, with an average of 0.22, meaning that 22 percent of neighbors folded into the same secondary structure as the wild-type [48]. Proteins can be even more robust than RNA. β-lactamase, an enzyme that endows bacteria with resistance to β-lactam antibiotics, can tolerate mutations in nearly 84 percent of its residues without impairing its function [49]. In the lac repressor, which is a DNA-binding protein that represses expression of genes involved in lactose metabolism, all the 301 amino acids that do not encode for the protein's DNA-binding domain can tolerate up to 70 percent of random substitutions, and the positions that bind to the DNA can tolerate up to 30 percent of random substitutions [50]. Isalan *et. al.* rewired bacterial regulatory

circuits by fusing promoter regions of a particular gene with the open reading frame of another gene, which resulted in 598 new regulatory interactions. Many of these fused promoter regions and genes were master regulators and sigma factors that control more than half of all *E. coli* genes[51]. Of the 598 clones, approximately 95 percent tolerated the rewiring, underscoring the robustness of the *E. coli* regulatory circuit [51].

All these examples from different biological systems reflect the organizing principle that makes *networks* of connected genotypes possible, namely that any one genotype has many neighboring genotypes with the same phenotype.

1.3.3 Innovation

In broad terms, biological innovation concerns the evolution of new traits. A single molecule, such as a protein may acquire a new enzymatic activity, or a combination of several enzymes may allow a bacterium to metabolize a non-natural chemical. Here, I am concerned with innovation that can be studied using genotype networks. As described earlier, an RNA molecule has many one-mutant neighbors with the same phenotype, which is a reflection of the mutational robustness of that genotype. However, its remaining one-mutant neighbors do not have the same phenotype. These neighbors constitute different novel phenotypes that lie in the vicinity of a specific genotype. For example, an RNA molecule with a defined secondary structure phenotype *P* and a mutational robustness of 0.25 means 75 percent of the molecule's neighbors have a secondary structure different from *P*. Specifically, 17 percent of the neighbors of random RNA molecules of length 30 nucleotides have the same secondary structure, while the remaining 83 percent fold into 32 different secondary structures [17]. A natural RNA molecule, the Hammerhead ribozyme of the peach latent mosaic viroid shows a similar diversity of phenotypes in its neighborhood [52].

The above example also suggests that higher mutational robustness of a genotype *G* means fewer genotypes with new phenotypes exist in the immediate neighborhood of genotype *G*. Consider an RNA genotype of length $L = 10$ nucleotides with a minimal robustness of zero, meaning that all 30 of its neighboring genotypes have a secondary

structure different form its own. On the other hand, consider a genotype of the same length (L = 10) on a genotype network of many genotypes, each of which has 20 neighbors with a novel phenotype. Starting from this genotype, one can access 20 different phenotypes in its own neighborhood, another 20 if it were to mutate to its neighboring genotype on the same network and so on [53]. However, it is possible that the neighborhood of each genotype in the genotype network consists of the same 20 novel phenotypes. In such as case, a genotype network would not facilitate innovation, because one could access only the 20 novel phenotypes. But if the neighborhoods of different genotypes contained different novel phenotypes, the existence of genotype networks becomes important to evolutionary innovations.

To understand if this is really the case, one can explore the genotype network while keeping track of the cumulative number of unique phenotypes that differ from our given phenotype. For example, to explore a genotype network of RNA with a given phenotype P, one chooses a genotype G at random, mutates a randomly chosen nucleotide in the RNA genotype and determines its secondary structure. If the new genotype has the same phenotype P, one then randomly chooses another nucleotide position to mutate. One thus repeats this process of carrying out single mutations on successive genotypes showing the same phenotype P while generating new genotypes G_D that differ in D nucleotides from G. Because G_D has the same phenotype P as genotype G, it lies on the same genotype network as G. Such exploration of the genotype network is known as a phenotype-preserving random walk (explained in greater detail below). Finally, one examines the one-mutant neighborhoods of genotypes G and G_D and counts the fraction of phenotypes in the neighborhood of G_D that do not occur in the one-mutant neighborhood of genotype G. A study of the Hammerhead ribozyme of the peach latent mosaic virus showed that genotypes at D = 2 contained more than 50 percent unique phenotypes in its neighborhood [52]. This fraction of unique phenotypes increased as the distance D increased [52]. Computational studies on random RNA molecules and experimental work on ribozymes show a similar pattern [53,54].

These observations are not limited to RNA molecules. For example, a relevant study analyzed the neighborhoods of 4950 pairs of metabolisms generated through long

random walks that preserved viability on glucose as a carbon source. While metabolisms generated through this procedure differed in the majority of their reactions ($D > 60$ percent), the phenotypes found in the neighborhoods of most metabolism pairs also differed to a great extent [55]. Gene regulatory circuits are not different. For instance, for circuits of 8 genes that differ in 6 percent of their regulatory interactions, 34 percent of new phenotypes occurring in the two-mutant neighborhood of two circuits are different [56]. This phenotype diversity increases with increasing differences in the regulatory interactions in these gene circuits [56]. All of the above evidence shows that genotype networks facilitate the exploration of novel phenotypes.

1.3.4 Latent phenotypes and exaptation

Exaptations or pre-adaptations are traits that arise as by-products of existing traits. These can be traits that were previously shaped by natural selection for a particular function and were later co-opted for another function. They can also be traits whose origin cannot be ascribed to the *direct* action of natural selection and had no use when the trait first arose [57]. A popular example of an exaptation is that of feathers, which originally served as thermal insulators and were later adopted for flight. Another example includes the Panda's thumb, which originated as a wrist-bone [57]. Both of these examples involve traits that were modified to assume the new function. Latent phenotypes, like exaptations, are also by-products of an existing trait, but may already demonstrate their novel phenotype. Such phenotypes can be later favored by natural selection to become exaptations [57,58]. A prominent example of a latent phenotype being selected for a new role involves the steroid hormone receptors, namely the mineralcorticoid and the glucocorticoid receptors. The mineralcorticoid receptor is activated by aldosterone and is present in tetrapods, while the glucocorticoid receptor is activated by cortisol, and is found in teleosts. Bridgham *et al*. inferred and resurrected their common ancestor, which was found to be sensitive to both cortisol and aldosterone [59]. However, tetrapods gained the ability to synthesize aldosterone millions of years after the divergence of tetrapods and teleosts. In other words, the ability of the mineralcorticoid receptor to bind aldosterone, a latent phenotype, was later exapted by tetrapods after they gained the ability to produce aldosterone [60]. An

additional example concerns the hedgehog gene regulatory circuit, which is involved in patterning the Drosophila wing blade [61], but was later exapted for the development of eyespots on butterfly wings [62], which play an important role in predator avoidance [63].

In the context of metabolism, latent phenotypes can also endow an organism with the ability to grow on new carbon sources. For example, in a laboratory evolution experiment, *Pseudomonas putida* S12 was evolved to efficiently utilize D-xylose as a carbon source. In addition to showing increased biomass yield on D-xylose, the evolved strains utilized L-arabinose efficiently, a carbon source that ancestral strains did not metabolize [64]. Another example includes promiscuous enzymes, which can catalyze reactions with a variety of substrates in vivo [65]. In a recent study, 37 percent of the 1081 metabolic enzymes in *E. coli* were found to be promiscuous [66].

For a genotype network with a given phenotype, genotypes within this network may have many such latent phenotypes. If the number of latent phenotypes shown by genotypes in a genotype network is high, it means that a genotype on such a network may have access to a large number of novel phenotypes. It also means that in suitable conditions, such latency can contribute to evolutionary innovation and adaptation. In Chapter 3, I analyze latent phenotypes shown by different metabolic genotypes with the same primary phenotype of being able to grow on a given carbon source.

1.3.5 Historical Contingency

A microorganism experiences mutations during evolution that change its genotype. Whether such mutations are deleterious, beneficial or neutral, that is, have no effect on the phenotype of the organism, may depend on the existing genotype. Mutations that occurred in the past may thus affect the effects of future mutations. This phenomenon is known as historical contingency. Such mutations can either result in small changes such as change in the activity of an enzyme, or may involve large changes such as loss-of-function mutations or gain of genetic material. To give an example, recent laboratory evolution experiments on multiple lines of *E. coli* showed that a specific genetic background was required in order to utilize citrate for survival

and growth [67]. A key mutation that enabled transport of citrate into the cell occurred in one of the *E. coli* populations. This mutation was then followed by other mutations that refined the population's ability to utilize citrate for the synthesis of all biomass precursors [67]. Thus, the mutation that enabled citrate transport was essential for the citrate utilization phenotype, and the mutations that arose later only refined this phenotype, but would not have been beneficial without the first mutation.

Whether the refining of an existing phenotype is possible may depend on the ability to start from a genotype with a given phenotype and reach other genotypes through a series of small genetic changes. Although up to this point I have treated genotype networks as if all genotypes that have a particular phenotype are connected through mutations the preserve the phenotype, this may not always be the case. Genotype networks can also be disconnected, wherein the genotype network fragments into multiple isolated subnetworks or *connected components* [68]. For example, genotype networks of RNA molecules can be fragmented, that is they form many such subnetworks [69,70], where it is not possible to connect a genotype in one component to a genotype in another component through a series of single nucleotide changes that preserve the phenotype. Similarly, a genotype network of model gene circuits that show dynamics similar to the yeast cell cycle fragments into 10^8 components [71]. One of the smaller components (comprising few genotypes) contains the wildtype circuit, meaning that the evolution of the wildtype circuit through small mutational steps is constrained only to other circuits in this component. On the other hand, evolution can explore a greater genotypic diversity in one of the larger components. These and other examples [72,73] indicate that historical contingency can be understood from the organization of genotype space. Because the extent of novel phenotypes that evolution may access depends strongly on a genotype's location in genotype space (see Section 1.3.3), fragmentation of genotype networks can also restrict the number of accessible novel phenotypes. In Chapter 4, I study the fragmentation of genotype networks of metabolism and show that historical contingency may not play an important role in the evolution of metabolic systems.

1.4 Computational modeling and simulation of metabolism

The computational study of metabolism requires models that are accurate in their biochemistry, can capture the functional states of a cell and allow computation of phenotypes, such as the viability of a cell or its growth rate in an environment. Different modeling formalisms exist, such as kinetic modeling, which has been used to model specific pathways [74], groups of pathways such as central carbon metabolism [75] and regulatory circuits [76,77]. While kinetic models describe dynamics of molecules and reactions in detail, they require large amounts of experimental data, such as kinetic parameters associated with an enzyme's activity. Conversely, constraint-based models of genome scale metabolism [8,41,78,79] require knowledge about specific enzymes and the reactions they catalyze. Secondly, while kinetic models are used to compute an explicit solution, constraint-based methods impose biologically motivated constraints that a metabolic system must satisfy. I briefly explain these different methods below.

1.4.1 Kinetic modeling

Kinetic modeling

A kinetic model of metabolic pathways uses ordinary differential equations to describe the dynamics of a metabolic system. These in turn can be derived from the theory of reaction kinetics. Briefly, the theory describes how substrate and product concentrations involved in a reaction change over time. For example, a basic form of a chemical reaction is the first order reaction

$A \xrightarrow{k} B$

wherein a chemical species A is converted to a product B at rate constant k. For an irreversible reaction, the rate of change of concentrations of the substrate $[A]$ and the product $[B]$ is given by

$$\frac{d[A]}{dt} = -k[A], \frac{d[B]}{dt} = k[A]$$

i.e. the rate of utilization of A and synthesis of B is linear in the concentration $[A]$. Metabolic reactions catalyzed by enzymes in most organisms are rarely that simple.

For example, a typical enzyme-catalyzed reaction involving substrate S, product P and free enzyme E has the following form when assuming irreversibility of the dissociation into the product:

$$E + S \underset{k_{-1}}{\overset{k_1}{\rightleftarrows}} ES \overset{k_2}{\rightarrow} E + P \qquad (1)$$

wherein ES denotes the substrate-bound form of the enzyme, also known as the enzyme-substrate complex. k_1 is the association rate constant of the enzyme-substrate complex, k_{-1} is the rate constant of the ES complex dissociating into free enzyme and substrate, while k_2 is the rate constant of the ES complex dissociating into free enzyme and product. Using these rate constants, one can then obtain the equation for the rate of change of the ES complex:

$$\frac{d[ES]}{dt} = k_1[E][S] - (k_{-1} + k_2)[ES]$$

This equation uses the steady-state approximation by Briggs and Haldane[80], which assumes that the concentration of the ES complex rapidly approaches steady-state, and does not change until most of the substrate has been consumed. This assumption corresponds to the equations

$$\frac{d[ES]}{dt} = 0$$

$$k_1[E][S] = (k_{-1} + k_2)[ES]$$

To determine the rate of product formation, given by $d[P]/dt = k_2[ES]$, or flux v of the reaction, one then introduces the total enzyme concentration E_t, which is equal to the total of the concentrations of ES and free enzyme E. Further rearranging of terms leads us to the equation

$$k_2[ES] = \frac{d[P]}{dt} = \frac{k_2[E_t][S]}{\frac{k_{-1} + k_2}{k_1} + [S]}$$

The final (modern) form of the Michaelis-Menten equation originally proposed in the year 1913 [81] is represented as

$$v = \frac{V_{max}[S]}{K_m + [S]} \qquad (2)$$

where $K_m = \frac{k_{-1}+k_2}{k_1}$ and $V_{max} = k_2[E_t]$

The above equation accounts for a reaction involving one substrate proceeding from the complex ES towards the formation of free enzyme E and product P in an irreversible fashion. However, biochemical reactions can be reversible, meaning that the catalytic reaction can proceed from substrate to product and vice-versa. The glycolytic enzyme glucosephosphate isomerase follows such a mechanism[82] and has the form

$$E + S \underset{k_{-1}}{\overset{k_1}{\rightleftarrows}} ES \underset{k_{-2}}{\overset{k_2}{\rightleftarrows}} E + P$$

Here, the rate of formation of the product P can be written as

$$\frac{d[P]}{dt} = v = k_2[ES] - k_{-2}[E][P]$$

In the absence of product, the above equation reverts to the irreversible catalytic reaction of equation (1), while in the absence of substrate, one can write equation (1) by replacing substrate S by product P. By comparing with equations (1) and (2), one can thus define parameters for the reverse reaction as

$$K_{mS} = \frac{k_{-1} + k_2}{k_1} ; Vf = k_2[E_t]$$

and

$$K_{mP} = \frac{k_{-1} + k_2}{k_{-2}} ; Vr = k_{-1}[E_t]$$

wherein Vf and Vr denote forward and reverse reaction velocities. The final form of a general reversible reaction can be thus written as

$$v = \frac{\frac{V_f S}{K_{mS}} - \frac{V_r P}{K_{mP}}}{1 + \frac{S}{K_{mS}} + \frac{P}{K_{mP}}} = \frac{V_f S K_{mP} - V_r P K_{mS}}{K_{mS} K_{mP} + S K_{mP} + P K_{mS}}$$

Another method of representing the reversible reaction is to understand that at equilibrium, the net rate of a reversible reaction is zero. This means that the forward and the reverse reactions exactly cancel each other. Thus, at equilibrium, the reversible Michaelis-Menten reaction reduces to

$$0 = \frac{V_f S_{eq}}{K_{mS}} - \frac{V_r P_{eq}}{K_{mP}}$$

wherein S_{eq} and P_{eq} denote concentrations of the substrate and the product at equilibrium. After rearrangement, the above equation allows us to compute the S_{eq} or the equilibrium constant of the reaction

$$K_{eq} = \frac{P_{eq}}{S_{eq}} = \frac{V_f K_{mP}}{V_r K_{mS}}$$

and

$$v = \frac{\frac{V_f}{K_{mS}} (S - \frac{P}{K_{eq}})}{1 + \frac{S}{K_{mS}} + \frac{P}{K_{mP}}}$$

This relationship is also known as the Haldane relationship[80,81].

Later work expanded the original form of the Michaelis-Menten equation to encompass more complex mechanisms [83]. In a series of papers, Cleland[84–86] defined the nomenclature, forms and equations of multi-substrate reactions that involve complex catalytic reactions wherein substrates can bind in a specific "sequential" order, in "ping-pong" fashion or a random order[84–86]. For example, reactions involving two substrates and two products are denoted as bi-bi reactions. If the substrates and products participating in such a reaction bind in no specific order, such a reaction is said to be random. If the binding of substrates *A* and *B* occurs in a specific order, the reaction is called ordered or sequential[86]. Such a reaction can be graphically represented as

$$E + A \underset{k_{-1}}{\overset{k_1}{\rightleftarrows}} EA + B \underset{k_{-2}}{\overset{k_2}{\rightleftarrows}} EAB \underset{k_{-3}}{\overset{k_3}{\rightleftarrows}} EPQ \underset{k_{-4}}{\overset{k_4}{\rightleftarrows}} EQ + P \underset{k_{-5}}{\overset{k_5}{\rightleftarrows}} E + Q$$

Here one assumes the formation of two central complexes *EAB* and *EPQ*. Some reactions can involve only one central complex *EAB*[86]. Not surprisingly, the rate equation of such a reaction is highly complicated. However, if one were to assume this reaction to be irreversible and to contain one central complex, the reaction can be represented as

$$E + A \underset{k_{-1}}{\overset{k_1}{\rightleftarrows}} EA + B \underset{k_{-2}}{\overset{k_2}{\rightleftarrows}} EAB \overset{k_3}{\rightarrow} E + P + Q$$

Note that binding of substrate *A* has to occur for substrate *B* to bind. The rate equation of such a reaction is represented[86] as

$$v = \frac{V_f\, AB}{K_{sA}K_{mB} + K_{mB}[A] + K_{mA}[B] + [A][B]}$$

Another important enzyme mechanism concerns cooperativity, which is found, for instance, in the binding of oxygen to the protein hemoglobin. Hemoglobin is an oligomer, meaning that an active protein consists of multiple protein subunits.

Cooperativity occurs when the binding of one oxygen molecule to a subunit alters the binding affinity of the remaining subunits to other oxygen molecules. This phenomenon can be described in the form of a Hill equation[87], given as

$$v = \frac{V_{max}[S]^n}{K_d + [S]^n}$$

where K_d denotes the dissociation constant of the substrate *S* to the protein, and *n* denotes the number of ligand molecules that bind to the protein. Enzymes that display cooperativity show sigmoidal reaction rates as one changes the substrate concentration as opposed to the hyperbolic shape displayed by enzymes following the Michaelis-Menten rule.

The above example deals with the binding of multiple copies of the same molecular species to the enzyme, a phenomenon also described as the homotropic effect[88]. However, many enzymes that exhibit cooperativity require more complex equations to describe their cooperativity. For example, an enzyme can bind to more than one ligand, where the binding of one ligand influences the binding of the other ligands positively or negatively (heterotrophy[88]). This change in the binding affinity of subsequent ligands does not occur through direct interference with the binding site, but may occur through a spatially different binding site and through conformational change in the protein. Proteins exhibiting such behavior are called allosteric proteins. A typical example involves the mammalian glycolytic enzyme phosphofructokinase, and its kinetic behavior follows the Monod-Wyman-Changeux or the concerted symmetry model[82,89,90]. An important purpose of glycolysis is to generate adenosine triphosphate (ATP), which acts as a feedback inhibitor of phosphofructokinase. The enzyme is also activated by its own substrate fructose-6-phosphate (F6P)[83]. Briefly, the concerted symmetry model requires that an enzyme's subunits exist in one of two different conformational states in the absence of a regulator molecule. One of these states is the “R” state, wherein a protein subunit has higher affinity for the substrate in comparison to its affinity for the ligand when in the other “T” state. A regulator molecule (ligand) can shift the conformation of all the subunits from one state to

another. Based on experimental observation, one can then compute the fraction of proteins in the R state, a state which is more conducive to catalysis[90].

The various catalytic mechanisms described above are few of the many enzyme mechanisms that are currently known. Experiments such as enzyme assays allow the determination of the enzyme mechanism and its associated kinetic parameters. One then uses such data to model each enzyme reaction. For modeling more than one such reaction, one writes differential equations for each of the reactants and finds the solution by numerical integration starting from a set of initial conditions. Although studies using such kinetic modeling abound, insufficient kinetic data limit its use for the modeling of genome-scale systems.

1.4.2 Flux balance analysis

Flux balance analysis (FBA) is a constraint-based method, wherein several constraints are applied to model a metabolic system based on its interaction with the environment. One such constraint concerns mass-balance, which assumes that all reactions internal to the system proceed at steady-state, meaning that all internal metabolites have to be synthesized and utilized at the same rate. The second constraint involves maximum and minimum bounds on reaction rates or fluxes in the metabolic system. Same principles hold for the import of an external metabolite into the system, which is defined by a nutrient medium[91,92]. Many solutions, that is the values of reaction fluxes can usually satisfy these constraints. To reduce this solution space, one specifies an optimization criterion. One such criterion can be the maximization of a desirable metabolic property, such as cell growth, for which different precursors of biomass (see Section 1.1) need to be produced in balanced amounts [93]. Such an optimization criterion assumes that cells maximize the molar conversion ratio or the molar yield of the precursors with respect to the provided nutrients, on the basis that cells use resources as efficiently as possible[94]. FBA then uses techniques such as linear programming to identify a set of flux values that satisfy the optimization criterion. Further details on FBA are provided in Chapters 2, 3 and 4.

Another constraint-based approach of studying metabolism is elementary flux modes. Elementary flux modes are minimal sets of biochemical reactions that can operate at steady-state between two external metabolites. They are minimal because complete inhibition of any one reaction within mode results in complete inhibition of the entire set of reactions. Thus, an elementary flux mode is nondecomposable and does not contain any alternative routes. As Schuster *et. al*. describe, any flux pattern in the metabolism can therefore be expressed as a linear combination of these modes[95]. The enumeration of elementary modes in a metabolism allows one to exhaustively study all possible metabolic routes within that metabolism.

1.4.3 Genome-scale metabolic reconstructions

Genome sequencing and bioinformatics has enabled the annotation and functional characterization for a large number of diverse microorganisms, which has led to the reconstruction of organism-specific models of metabolism. Such models have been constructed for more than 50 prokaryotic microorganisms alone [96], each involving 500-1500 reactions.

Genome-scale metabolic reconstructions are based on the sequenced genome of the microorganism and curated manually using information from several metabolic databases such as KEGG [39,97,98], BRENDA [99], as well as primary literature. This process results in a list of metabolic reactions taking place in an organism, along with accurate stoichiometric information about the reactants participating in each reaction. In addition, one also needs knowledge about the biomass precursors that are required for the survival and growth of the cell. For well-studied organisms such as *E. coli*, the proportions of biomass precursors are known[4,6,8]. One also needs to understand the energy requirements of growth and non-growth-associated maintenance for an organism [100]. Although reconstructions include reactions that use ATP as a substrate, other processes that require energy, such as the assembly of biopolymers (DNA, proteins) are not explicitly modeled in the reconstruction. These requirements are called growth-associated maintenance. Non-growth-associated maintenance requirements involve maintenance of the proton gradient against the leakage of protons across the cell membrane [8]. These energy requirements are often organism-

specific and can be found in the literature or are determined experimentally [100–102]. Once the above pre-requisites are included in the reconstruction, one can use constraint-based methods such as flux balance analysis to understand the metabolic behavior of the model in different growth environments. These growth environments include the rate of uptake for different nutrients that serve as sources for carbon, nitrogen, sulfur, phosphorous, to name a few. One then tests model predictions with laboratory experiments, such as the identification of reactions that are essential for growth of the microorganism. Discrepancies in the model are then resolved by an iterative cycle of comparison with experimental data towards increasing model accuracy [8,96,103–105].

1.4.4 Markov chain Monte Carlo (MCMC) sampling

Sampling of the optimal flux space. Although constraint-based methods such as FBA may correctly predict the rate of biomass synthesis, many different solutions (flux values) can satisfy this condition. For example, if two alternate pathways can synthesize a given metabolite at a maximal rate, any one of two pathways could be maximally active while the other pathway has no flux, or both of the pathways could be active such that maximal synthesis of the given metabolite is achieved. In other words, a metabolic system may contain multiple optimal solutions, a fact that gives rise to the phenomenon of an "optimal flux space" [106]. It is instructive to study which multiple optima is the most probable. Markov chain Monte Carlo (MCMC) sampling achieves this through randomly sampling the flux space in a uniform manner.

One of the earliest studies that used such sampling analyzed the metabolism of the red blood cell and computed the distribution of flux values for each of the reactions in the metabolic system. They found that certain reactions are constrained to a narrow distribution of flux values, while others are not [107]. A later study employed a more refined approach that could sample larger metabolisms [108] and used it to analyze correlations of fluxes in the metabolism of the red blood cell. It found that many reactions are correlated in the distribution of their flux values. In addition, the study modeled the systemic effects of a reduction in the maximal reaction rate (V_{max}) to simulate known enzymopathies. It showed that not only the range of flux values change throughout the system, but also their correlations with each other [108]. Further

refined versions of MCMC sampling have enabled sampling of the flux space in genome-scale metabolisms, leading to the elucidation of a common core network of reactions with high rates in different environments [109] and in the metabolisms of several organisms [110].

Sampling of the genotype network of metabolisms. The known "universe" of metabolic reactions contains more than 5000 reactions. Assuming the presence or absence of reactions, the space of possible metabolisms contains more than 25000 members. While all genotypes in such a vast space cannot be enumerated, MCMC sampling allows a systematic exploration of this space by generating a large number of metabolisms that have the same phenotype, that is, they are viable on a given carbon source. Briefly, starting from a metabolic genotype, one adds a randomly chosen reaction from the reaction universe, followed by the deletion of another randomly chosen reaction from the genotype[111]. FBA is then used to determine whether the reaction deletion abolished viability of the metabolism on the given carbon source. The deletion step is accepted only if the metabolism remains viable after deletion. If not, another reaction is randomly chosen for removal [111]. This sequence of reaction addition and deletion is called a reaction swap and the process of carrying out such successive reaction swaps is called a random walk. Metabolisms generated after such long random walks are called random viable metabolisms. Previous studies have used MCMC sampling to study the evolution of microbial metabolism [55,111,112], as I do in my analyses here. Further details on MCMC sampling of random viable metabolisms are provided in Chapters 2, 3 and 4.

1.5 Applications of genome scale models

Genome-scale metabolic models have been used in many different areas such as evolutionary biology, biotechnology and medicine. Next, I review some of these.

1.5.1 Evolutionary applications

Gene dispensability and mutational robustness. One of the most important discoveries in microbiology in recent years has been the proportion of genes that appear to be

non-essential for the growth and survival of an organism. Large-scale gene deletion studies have demonstrated that approximately 80 percent of protein-coding genes of organisms such as *E. coli* and *S. cerevisiae* are not essential for viability in laboratory conditions[113,114]. These findings pose questions about the evolution of metabolic systems to tolerate such mutations. Computational and experimental studies have shown that many (37 to 68 percent) of the apparently dispensable genes are actually essential in other different environmental conditions, while other non-essential genes are involved in alternate metabolic routes that allow the rerouting of fluxes [115].

The dispensability of metabolic reactions also speaks to the mutational robustness of metabolism, defined as the proportion of non-essential reactions that do not abolish viability when deleted. Although free-living organisms such as *E. coli* have high mutational robustness, one can quantify how different is this mutational robustness in comparison to a typical metabolism. Towards answering this question, Samal *et. al.* used Markov Chain Monte Carlo methods to sample ensembles of random metabolisms viable on specific environments [111] (see Section 1.4.4). They found *E. coli* metabolism to be significantly more robust in comparison to random metabolisms viable in different environments. Furthermore, mutational robustness of random metabolisms in different environments was correlated, meaning that an increase in robustness in an environment may increase robustness in another environment as well [111], meaning that reactions that contribute to higher mutational robustness may be shared between different environments.

Sequence and genome evolution. The evolution of metabolism is not restricted to loss and gain of genes, but can also occur through point mutations in genes. Such mutations can directly affect enzyme activity or change the expression of enzymes. Using a metabolic model of the yeast *S. cerevisiae*, a study showed that enzymes that catalyze reactions with high metabolic fluxes are also especially sensitive to amino acid changes [116]. Furthermore, central and highly connected enzymes, i.e. enzymes that share their reactants with many other enzymes, tolerate fewer changes [116]. In a similar vein, another study quantified the range of flux values for different reactions that result in optimal and near-optimal growth, wherein this range of flux values can be thought of reflecting possible expression divergence of genes. This study found

that genes which display greater expression divergence across different strains of yeast also have a broader range of metabolic fluxes [117]. Together, these studies show that the evolution of gene sequence and expression is linked to the metabolic role of a gene in the metabolism of an organism.

Given that horizontal gene transfer is one of the most important drivers of metabolic evolution, understanding the metabolic roles of horizontally transferred genes can be instructive towards understanding the evolution of metabolism. Pál *et. al.* identified genes that have been recently horizontally transferred in to the *E. coli* genome and found them preferentially at the periphery of metabolism, wherein they facilitate processes such as transport of nutrients and their breakdown and assimilation in the central carbon metabolism [118]. Computational modeling using FBA showed that the newly added genetic material is responsible for reactions that allowed *E. coli* to adapt to new environmental conditions [118].

Endosymbiotic bacteria such as *B. aphidicola* have undergone a loss of 75 percent of their protein-coding genes since their split from the *E. coli* lineage 200 million years ago [119]. Endosymbionts get access to many nutrients from their host, resulting in relaxed selection pressure and subsequent accumulation of deleterious mutations in many metabolic and other genes [119]. However, different *B. aphidicola* strains show variation in gene content. A study used the genome-scale reconstruction of *E. coli* metabolism to computationally model successive gene loss events and correctly predicted 80 percent of the gene loss events that *B. aphidicola* experienced [120].

Adaptive evolution in the laboratory. A major success story of genome-scale reconstructions involves the ability to predict phenotypic outcomes of laboratory evolution. In a pioneering study, Palsson and group combined laboratory evolution and computational modeling to study growth evolution of the *E. coli* K12 strain with glycerol as the sole carbon source [121]. Wild-type *E. coli* has a glycerol utilization pathway, but it grows suboptimally on this medium [121]. However, after 700 generations of evolution a strain emerged that showed a dramatic increase in growth. The nutrient uptake and growth rate of the evolved strain was in good agreement with the predicted optimal metabolic state [121]. In the same vein, end-point phenotypes of

the evolution experiment such as nutrient uptake, by-product secretion patterns and growth rates resulting from the adaptive evolution of *E. coli* on different sole carbon sources can be accurately predicted by computational modeling [122]. Lastly, the Palsson and co-workers created knockout strains of various metabolic enzymes of *E. coli*. While the growth rate immediately after deletion could not be predicted by modeling, their end-point growth rates after adaptive evolution could be predicted with a success rate of 78 percent [123], highlighting the efficacy of flux balance analysis in predicting endpoints of evolution. Other constraint-based methods such as ROOM (regulatory on-off minimization) and MOMA (minimization of metabolic adjustment) have met with reasonable success in predicting the immediate suboptimal phenotype of deletion strains not subjected to adaptive evolution [124,125].

1.5.2 Biotechnology and Medicine

Metabolic engineering. Since genome-scale metabolic reconstructions allow accurate prediction of cellular phenotypes after genetic and environmental perturbations, they have also been used for biotechnological applications and metabolic engineering. As a prominent example, *E. coli* metabolic reconstructions have been used to identify metabolic engineering strategies for the overproduction of many compounds. Examples include the overproduction of amino acids such as L-valine[126] and L-threonine [127], as well as other industrially important metabolites such as malate[128], lactate [129], succinate [130,131], and lycopene [132,133].

Secretion of important intermediate metabolites diverts a significant amount of nutrients from the synthesis of biomass precursors and thus conflicts with maximal growth. Furthermore, metabolism is known to be highly reticulate; many biochemical pathways are linked through shared reactants and biochemical feedback systems [134,135]. One thus needs to redesign bacterial metabolism through strategies such as gene knockouts and introduction of genes from other organisms to guarantee overproduction of compounds of interest [136]. Several approaches based on constraint-based modeling have been developed towards the optimization of conflicting goals of maximizing metabolite production and biomass synthesis. The earliest of these, OptKnock predicts double, triple and quadruple gene knockouts that force metabolic

flux towards overproduction of a desired metabolite [137]. OptKnock has been used successfully to enable overproduction of lactate through triple and quadruple gene knockouts [129]. OptKnock was also used to correctly predict gene knockouts that lead to overproduction of 1,4-butanediol, a commercially important compound [138]. Another approach called OptStrain was used to correctly predict the addition of non-native reactions to the *E. coli* metabolism in order to synthesize vanillin, a compound not produced by wild-type *E. coli* [139]. Other efficient computational approaches have been developed and used for the rational design of microbial metabolism [140–144].

Medicine. The development of antibiotics is a time- and money-intensive endeavor. The most important criterion that an ideal drug target should fulfill is that the target protein or enzyme should be essential for the survival of the microorganism. The identification of essential metabolic processes through computational approaches helps to expedite this process. Numerous metabolic reconstructions of pathogens have been developed with the primary aim of identifying drug targets that inhibit cellular function. Here are a few examples.

Mycobacterium tuberculosis is an important cause of tuberculosis, especially in developing countries. One of the earliest studies involving *M. tuberculosis* analyzed the synthesis of mycolate through mycolic acid pathway. Mycolate is a long fatty acid found in the cell wall of *M. tuberculosis* and is important for the organism's survival, growth and pathogenicity. The study predicted new drug targets based on the computational identification of gene knockouts that abolished the synthesis of mycolate [145]. A later study used a reconstruction of *M. tuberculosis*, together with flux balance analysis and gene expression data to correctly predict the effects of different drugs and drug combinations on *M. tuberculosis* growth and survival across several nutrient conditions [146]. Another study identified "hard-coupled" reaction sets in the metabolic model of *M. tuberculosis*, wherein groups of connected reactions operate in unison because of mass conservation constraints [147]. Deletion of any single reaction in a set eliminates the functionality of the entire set [147], resulting in the identification of new drug targets. Similar studies using the metabolic reconstruction of *Staphylococcus aureus* have been carried out to identify essential metabolites and reactions for *S. aureus* survival [148,149]. One study used sequence homology data to

reconstruct metabolic models of 13 *S. aureus* strains [150]. It predicted essential reactions catalyzed by 70 different enzymes and 54 enzyme pairs, of which reactions catalyzed by 44 single enzymes and 10 enzyme pairs were common to all 13 *S. aureus* strains. Many of these reactions were previously unidentified. Many other pathogens have been studied using metabolic reconstruction, including *H. influenza* [151], *P. aeruginosa* [152], *H. pylori* [44,148,153], *S. typhimirium* [78,154], *N. meningitides* [155], and *M. genitalium* [156]. The case studies discussed here are testimony to the use and predictive power of genome-scale reconstructions.

1.6 Thesis outline

Among the many different areas of evolutionary biology, this thesis focuses on the evolution of microbial metabolism. This has become feasible thanks to (i) our current knowledge of the universe of reactions, which forms a vast space of all possible metabolisms (ii) computational constraint-based approaches that allow us to model and simulate metabolic systems, and (iii) approaches that enable sampling of large metabolic spaces, such as Markov chain Monte Carlo (MCMC) sampling. These developments are important because the study of metabolism is contingent on being able to analyze ensembles of metabolisms that have a biologically relevant phenotype, such as viability on a given carbon source. MCMC sampling along with flux balance analysis allows one to sample such random metabolisms that have the same phenotype.

One way of studying the evolution of metabolism is to study essential reactions, which are metabolic reactions that are necessary for the survival and growth of an organism. In **chapter 2**, I determine the essentiality of each reaction in the universe of reactions for growth on a given carbon source. In doing that, I find that any metabolic genotype contains a core of reactions that cannot be replaced by any other reactions. Whether reaction essentiality influences the maintenance of enzyme-coding genes in prokaryotic genomes is discussed. The observations in this study are also relevant for the identification of drug targets against infectious diseases.

Chapter 3 focuses on the origin of latent metabolic traits that are non-adaptive in origin. These traits concern the ability to grow on novel carbon sources. Specifically, can metabolic genotypes with a given phenotype exhibit latent phenotypes? How are these latent traits influenced by the primary phenotype? I address these and other related questions in chapter 3. This study suggests that latent metabolic phenotypes may be more ubiquitous than is currently appreciated.

Genotype networks of biological systems can be fragmented, underscoring the importance of historical contingency in the evolution of metabolic properties such as mutational robustness. **Chapter 4** asks if such historical contingency plays a role in

the evolution of metabolism. I study genotype networks of simple and genome scale metabolisms and determine their connectedness. In doing that, I find that genotype networks of only the smallest metabolic systems are disconnected. Furthermore, I also discuss possible reasons for the disconnectedness in these genotype networks. My observations suggest historical contingency may not greatly influence the evolution of metabolic systems in general.

1.7 References

1. Canfield, D. E. & Des Marais, D. J. Biogeochemical cycles of carbon, sulfur, and free oxygen in a microbial mat. *Geochim. Cosmochim. Acta* **57,** 3971–84 (1993).

2. Falkowski, P. G. Evolution of the nitrogen cycle and its influence on the biological sequestration of CO2 in the ocean. *Nature* **387,** 272–275 (1997).

3. Amann, R. I., Ludwig, W. & Schleifer, K. H. Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiol. Rev.* **59,** 143–69 (1995).

4. Neidhardt, F. *Escherichia coli and salmonella : cellular and molecular biology*. (1996).

5. Noor, E., Eden, E., Milo, R. & Alon, U. Central carbon metabolism as a minimal biochemical walk between precursors for biomass and energy. *Mol. Cell* **39,** 809–20 (2010).

6. Smith, C. A. Physiology of the Bacterial Cell. A Molecular Approach. *Biochem. Educ.* **20,** 124–125 (1992).

7. Palsson, B. O. *Systems Biology: Properties of Reconstructed Networks*. (Cambridge University Press, 2006).

8. Feist, A. M. *et al.* A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol. Syst. Biol.* **3,** 121 (2007).

9. Csete, M. & Doyle, J. Bow ties, metabolism and disease. *Trends Biotechnol.* **22,** 446–50 (2004).

10. Sauer, U. Metabolic networks in motion: 13C-based flux analysis. *Mol. Syst. Biol.* **2,** 62 (2006).

11. Ma, H.-W. & Zeng, A.-P. The connectivity structure, giant strong component and centrality of metabolic networks. *Bioinformatics* **19,** 1423–30 (2003).

12. Blaby, I. K. *et al.* Experimental evolution of a facultative thermophile from a mesophilic ancestor. *Appl. Environ. Microbiol.* **78,** 144–55 (2012).

13. Applebee, M. K., Herrgård, M. J. & Palsson, B. Ø. Impact of individual mutations on increased fitness in adaptively evolved strains of Escherichia coli. *J. Bacteriol.* **190,** 5087–94 (2008).

14. Lee, D.-H. & Palsson, B. Ø. Adaptive evolution of Escherichia coli K-12 MG1655 during growth on a Nonnative carbon source, L-1,2-propanediol. *Appl. Environ. Microbiol.* **76,** 4158–68 (2010).

15. Charusanti, P. *et al.* Genetic basis of growth adaptation of Escherichia coli after deletion of pgi, a major metabolic gene. *PLoS Genet.* **6,** e1001186 (2010).

16. Blattner, F. R. *et al.* The complete genome sequence of Escherichia coli K-12. *Science (80-.).* **277,** 1453–62 (1997).

17. Wagner, A. *The origins of evolutionary innovations*. in press (Oxford University Press, USA, 2011).

18. Ochman, H. & Jones, I. B. Evolutionary dynamics of full genome content in Escherichia coli. *EMBO J.* **19,** 6637–43 (2000).

19. Wagner, A. Evolutionary constraints permeate large metabolic networks. *BMC Evol. Biol.* **9,** 231 (2009).

20. Kloesges, T., Popa, O., Martin, W. & Dagan, T. Networks of gene sharing among 329 proteobacterial genomes reveal differences in lateral gene transfer frequency at different phylogenetic depths. *Mol. Biol. Evol.* **28,** 1057–74 (2011).

21. Lerat, E., Daubin, V., Ochman, H. & Moran, N. A. Evolutionary origins of genomic repertoires in bacteria. *PLoS Biol.* **3,** e130 (2005).

22. Ma, Y.-F. *et al.* Nucleotide sequence of plasmid pCNB1 from comamonas strain CNB-1 reveals novel genetic organization and evolution for 4-chloronitrobenzene degradation. *Appl. Environ. Microbiol.* **73,** 4477–83 (2007).

23. Williams, P. A. & Sayers, J. R. The evolution of pathways for aromatic hydrocarbon oxidation in Pseudomonas. *Biodegradation* **5,** 195–217 (1994).

24. Hehemann, J.-H. *et al.* Transfer of carbohydrate-active enzymes from marine bacteria to Japanese gut microbiota. *Nature* **464,** 908–912 (2010).

25. Roberts, M. C. Update on acquired tetracycline resistance genes. *FEMS Microbiol. Lett.* **245,** 195–203 (2005).

26. Roberts, M. C. Update on macrolide-lincosamide-streptogramin, ketolide, and oxazolidinone resistance genes. *FEMS Microbiol. Lett.* **282,** 147–59 (2008).

27. Pantosti, A., Sanchini, A. & Monaco, M. Mechanisms of antibiotic resistance in Staphylococcus aureus. *Future Microbiol.* **2,** 323–34 (2007).

28. Dubnau, D. DNA uptake in bacteria. *Annu. Rev. Microbiol.* **53,** 217–44 (1999).

29. Jonas, D. A. *et al.* Safety considerations of DNA in food. *Ann. Nutr. Metab.* **45,** 235–54 (2001).

30. Lorenz, M. G. & Wackernagel, W. Bacterial gene transfer by natural genetic transformation in the environment. *Microbiol. Rev.* **58,** 563–602 (1994).

31. Marraffini, L. A. & Sontheimer, E. J. CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. *Nat. Rev. Genet.* **11,** 181–90 (2010).

32. Heinemann, J. A. & Sprague, G. F. Bacterial conjugative plasmids mobilize DNA transfer between bacteria and yeast. *Nature* **340,** 205–9 (1989).

33. Buchanan-Wollaston, V., Passiatore, J. E. & Cannon, F. The mob and oriT mobilization functions of a bacterial plasmid promote its transfer to plants. *Nature* **328,** 172–175 (1987).

34. Majewski, J. Sexual isolation in bacteria. *FEMS Microbiol. Lett.* **199,** 161–9 (2001).

35. Dorman, C. J. Regulatory integration of horizontally-transferred genes in bacteria. *Front. Biosci. (Landmark Ed.* **14,** 4103–12 (2009).

36. Sorek, R. *et al.* Genome-wide experimental determination of barriers to horizontal gene transfer. *Science* **318,** 1449–52 (2007).

37. Schuster, P., Fontana, W., Stadler, P. F. & Hofacker, I. L. From sequences to shapes and back: a case study in RNA secondary structures. *Proc. Biol. Sci.* **255,** 279–84 (1994).

38. Orengo, C. A., Jones, D. T. & Thornton, J. M. Protein superfamilies and domain superfolds. *Nature* **372,** 631–4 (1994).

39. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28,** 27–30 (2000).

40. Goto, S., Nishioka, T. & Kanehisa, M. LIGAND: chemical database of enzyme reactions. *Nucleic Acids Res.* **28,** 380–2 (2000).

41. Baumler, D. J., Peplinski, R. G., Reed, J. L., Glasner, J. D. & Perna, N. T. The evolution of metabolic networks of E. coli. *BMC Syst. Biol.* **5,** 182 (2011).

42. Raghunathan, A., Reed, J., Shin, S., Palsson, B. & Daefler, S. Constraint-based analysis of metabolic capacity of Salmonella typhimurium during host-pathogen interaction. *BMC Syst. Biol.* **3,** 38 (2009).

43. Oh, Y.-K., Palsson, B. O., Park, S. M., Schilling, C. H. & Mahadevan, R. Genome-scale reconstruction of metabolic network in Bacillus subtilis based on high-throughput phenotyping and gene essentiality data. *J. Biol. Chem.* **282,** 28791–9 (2007).

44. Thiele, I., Vo, T. D., Price, N. D. & Palsson, B. Ø. Expanded metabolic reconstruction of Helicobacter pylori (iIT341 GSM/GPR): an in silico genome-scale characterization of single- and double-deletion mutants. *J. Bacteriol.* **187,** 5818–30 (2005).

45. Schuster, P. & Fontana, W. Chance and necessity in evolution: lessons from RNA. *Phys. D Nonlinear Phenom.* **133,** 427–452 (1999).

46. Reidhaar-Olson, J. F. & Sauer, R. T. Functionally acceptable substitutions in two alpha-helical regions of lambda repressor. *Proteins* **7,** 306–16 (1990).

47. Jörg, T., Martin, O. C. & Wagner, A. Neutral network sizes of biological RNA molecules can be computed and are not atypically small. *BMC Bioinformatics* **9,** 464 (2008).

48. Sanjuán, R., Forment, J. & Elena, S. F. In silico predicted robustness of viroids RNA secondary structures. I. The effect of single mutations. *Mol. Biol. Evol.* **23,** 1427–36 (2006).

49. Huang, W., Petrosino, J., Hirsch, M., Shenkin, P. S. & Palzkill, T. Amino acid sequence determinants of beta-lactamase structure and activity. *J. Mol. Biol.* **258,** 688–703 (1996).

50. Kleina, L. G. & Miller, J. H. Genetic studies of the lac repressor. XIII. Extensive amino acid replacements generated by the use of natural and synthetic nonsense suppressors. *J. Mol. Biol.* **212,** 295–318 (1990).

51. Isalan, M. *et al.* Evolvability and hierarchy in rewired bacterial gene networks. *Nature* **452,** 840–5 (2008).

52. Sumedha, Martin, O. C. & Wagner, A. New structural variation in evolutionary searches of RNA neutral networks. *Biosystems.* **90,** 475–85

53. Wagner, A. Robustness and evolvability: a paradox resolved. *Proc. Biol. Sci.* **275,** 91–100 (2008).

54. Schultes, E. A. & Bartel, D. P. One sequence, two ribozymes: implications for the emergence of new ribozyme folds. *Science* **289,** 448–52 (2000).

55. Matias Rodrigues, J. F. & Wagner, A. Evolutionary plasticity and innovations in complex metabolic reaction networks. *PLoS Comput. Biol.* **5,** e1000613 (2009).

56. Ciliberti, S., Martin, O. C. & Wagner, A. Innovation and robustness in complex regulatory gene networks. *Proc. Natl. Acad. Sci. U. S. A.* **104,** 13591–6 (2007).

57. Gould, S. J. & Vrba, E. S. Exaptation-a missing term in the science of form. *Paleobiology* **8,** 4–15 (1982).

58. True, J. R. & Carroll, S. B. Gene co-option in physiological and morphological evolution. *Annu. Rev. Cell Dev. Biol.* **18,** 53–80 (2002).

59. Bridgham, J. T., Carroll, S. M. & Thornton, J. W. Evolution of hormone-receptor complexity by molecular exploitation. *Science* **312,** 97–101 (2006).

60. Conant, G. C. & Wolfe, K. H. Turning a hobby into a job: how duplicated genes find new functions. *Nat. Rev. Genet.* **9,** 938–50 (2008).

61. Brook, W. J., Diaz-Benjumea, F. J. & Cohen, S. M. Organizing spatial pattern in limb development. *Annu. Rev. Cell Dev. Biol.* **12,** 161–80 (1996).

62. Keys, D. N. *et al.* Recruitment of a hedgehog regulatory circuit in butterfly eyespot evolution. *Science* **283,** 532–4 (1999).

63. BRAKEFIELD, P. M. & REITSMA, N. Phenotypic plasticity, seasonal climate and the population biology of Bicyclus butterflies (Satyridae) in Malawi. *Ecol. Entomol.* **16,** 291–303 (1991).

64. Meijnen, J.-P., de Winde, J. H. & Ruijssenaars, H. J. Engineering Pseudomonas putida S12 for efficient utilization of D-xylose and L-arabinose. *Appl. Environ. Microbiol.* **74,** 5031–7 (2008).

65. Khersonsky, O. & Tawfik, D. S. Enzyme promiscuity: a mechanistic and evolutionary perspective. *Annu. Rev. Biochem.* **79,** 471–505 (2010).

66. Nam, H. *et al.* Network context and selection in the evolution to enzyme specificity. *Science* **337,** 1101–4 (2012).

67. Blount, Z. D., Borland, C. Z. & Lenski, R. E. Historical contingency and the evolution of a key innovation in an experimental population of Escherichia coli. *Proc. Natl. Acad. Sci. U. S. A.* **105,** 7899–906 (2008).

68. Newman, M. *Networks : an introduction*. (Oxford University Press, 2010).

69. Schaper, S., Johnston, I. G. & Louis, A. A. Epistasis can lead to fragmented neutral spaces and contingency in evolution. *Proc. Biol. Sci.* **279,** 1777–83 (2012).

70. Aguirre, J., Buldú, J. M., Stich, M. & Manrubia, S. C. Topological structure of the space of phenotypes: the case of RNA neutral networks. *PLoS One* **6,** e26324 (2011).

71. Boldhaus, G. & Klemm, K. Regulatory networks and connected components of the neutral space. *Eur. Phys. J. B* **77,** 233–237 (2010).

72. Payne, J. L. & Wagner, A. Constraint and contingency in multifunctional gene regulatory circuits. *PLoS Comput. Biol.* **9,** e1003071 (2013).

73. Ciliberti, S., Martin, O. C. & Wagner, A. Robustness can evolve gradually in complex regulatory gene networks with varying topology. *PLoS Comput. Biol.* **3,** e15 (2007).

74. Teusink, B., Walsh, M. C., van Dam, K. & Westerhoff, H. V. The danger of metabolic pathways with turbo design. *Trends Biochem. Sci.* **23,** 162–9 (1998).

75. Barve, A., Gupta, A. & Solapure, S. M. A kinetic platform for in silico modeling of the metabolic dynamics in Escherichia coli. *Adv. Appl. Bioinforma. Chem.* **2010,** 97–110 (2010).

76. Chen, K. C. *et al.* Integrative analysis of cell cycle control in budding yeast. *Mol. Biol. Cell* **15,** 3841–62 (2004).

77. Sha, W. *et al.* Hysteresis drives cell-cycle transitions in Xenopus laevis egg extracts. *Proc. Natl. Acad. Sci. U. S. A.* **100,** 975–80 (2003).

78. Raghunathan, A., Reed, J., Shin, S., Palsson, B. & Daefler, S. Constraint-based analysis of metabolic capacity of Salmonella typhimurium during host-pathogen interaction. *BMC Syst. Biol.* **3,** 38 (2009).

79. Reed, J. L., Vo, T. D., Schilling, C. H. & Palsson, B. O. An expanded genome-scale model of Escherichia coli K-12 (iJR904 GSM/GPR). *Genome Biol.* **4,** R54 (2003).

80. Briggs, G. E. & Haldane, J. B. A Note on the Kinetics of Enzyme Action. *Biochem. J.* **19,** 338–9 (1925).

81. Michaelis, L., Menten, M. L., Johnson, K. A. & Goody, R. S. The original Michaelis constant: translation of the 1913 Michaelis-Menten paper. *Biochemistry* **50,** 8264–9 (2011).

82. Mulquiney, P. J. & Kuchel, P. W. Model of 2,3-bisphosphoglycerate metabolism in the human erythrocyte based on detailed enzyme kinetic equations: equations and parameter refinement. *Biochem. J.* **342 Pt 3,** 581–96 (1999).

83. Segel, I. H. *Enzyme Kinetics: Behavior and Analysis of Rapid Equilibrium and Steady-State Enzyme Systems*. 957 (Wiley, 1993).

84. Cleland, W. . The kinetics of enzyme-catalyzed reactions with two or more substrates or products. III. Prediction of initial velocity and inhibition patterns by inspection. *Biochim. Biophys. Acta - Spec. Sect. Enzymol. Subj.* **67,** 188–196 (1963).

85. Cleland, W. . The kinetics of enzyme-catalyzed reactions with two or more substrates or products. II. Inhibition: nomenclature and theory. *Biochim. Biophys. Acta - Spec. Sect. Enzymol. Subj.* **67,** 173–187 (1963).

86. Cleland, W. . The kinetics of enzyme-catalyzed reactions with two or more substrates or products. I. Nomenclature and rate equations. *Biochim. Biophys. Acta - Spec. Sect. Enzymol. Subj.* **67,** 104–137 (1963).

87. Edelstein, S. J. Cooperative interactions of hemoglobin. *Annu. Rev. Biochem.* **44,** 209–32 (1975).

88. Tsuneshige, A., Park, S. & Yonetani, T. Heterotropic effectors control the hemoglobin function by interacting with its T and R states--a new view on the principle of allostery. *Biophys. Chem.* **98,** 49–63 (2002).

89. Roy, H., Diwan, J., D. Segel, L. & Segel, I. H. Computer-assisted simulations of phosphofructokinase-1 kinetics using simplified velocity equations. *Biochem. Mol. Biol. Educ.* **29,** 3–9 (2001).

90. Changeux, J.-P. 50 years of allosteric interactions: the twists and turns of the models. *Nat. Rev. Mol. Cell Biol.* **14,** 819–29 (2013).

91. Kauffman, K. J., Prakash, P. & Edwards, J. S. Advances in flux balance analysis. *Curr. Opin. Biotechnol.* **14,** 491–6 (2003).

92. Price, N. D., Reed, J. L. & Palsson, B. Ø. Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nat. Rev. Microbiol.* **2,** 886–97 (2004).

93. Schuetz, R., Kuepfer, L. & Sauer, U. Systematic evaluation of objective functions for predicting intracellular fluxes in Escherichia coli. *Mol. Syst. Biol.* **3,** 119 (2007).

94. Schuster, S., Pfeiffer, T. & Fell, D. A. Is maximization of molar yield in metabolic networks favoured by evolution? *J. Theor. Biol.* **252,** 497–504 (2008).

95. Schuster, S., Fell, D. A. & Dandekar, T. A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nat. Biotechnol.* **18,** 326–32 (2000).

96. Feist, A. M., Herrgård, M. J., Thiele, I., Reed, J. L. & Palsson, B. Ø. Reconstruction of biochemical networks in microorganisms. *Nat. Rev. Microbiol.* **7,** 129–43 (2009).

97. Goto, S., Okuno, Y., Hattori, M., Nishioka, T. & Kanehisa, M. LIGAND: database of chemical compounds and reactions in biological pathways. *Nucleic Acids Res.* **30,** 402–4 (2002).

98. Kanehisa, M. *et al.* From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.* **34,** D354–7 (2006).

99. Schomburg, I. *et al.* BRENDA in 2013: integrated reactions, kinetic data, enzyme function data, improved disease classification: new options and contents in BRENDA. *Nucleic Acids Res.* **41,** D764–72 (2013).

100. Pirt, S. J. The maintenance energy of bacteria in growing cultures. *Proc. R. Soc. Lond. B. Biol. Sci.* **163,** 224–31 (1965).

101. Taymaz-Nikerel, H., Borujeni, A. E., Verheijen, P. J. T., Heijnen, J. J. & van Gulik, W. M. Genome-derived minimal metabolic models for Escherichia coli MG1655 with estimated in vivo respiratory ATP stoichiometry. *Biotechnol. Bioeng.* **107,** 369–81 (2010).

102. Baart, G. J. E. *et al.* Modeling Neisseria meningitidis B metabolism at different specific growth rates. *Biotechnol. Bioeng.* **101,** 1022–35 (2008).

103. Baart, G. J. E. & Martens, D. E. Genome-scale metabolic models: reconstruction and analysis. *Methods Mol. Biol.* **799,** 107–26 (2012).

104. Orth, J. D. *et al.* A comprehensive genome-scale reconstruction of Escherichia coli metabolism--2011. *Mol. Syst. Biol.* **7,** 535 (2011).

105. Henry, C. S. *et al.* High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nat. Biotechnol.* **28,** 977–82 (2010).

106. Varma, A. & Palsson, B. O. Metabolic capabilities of Escherichia coli: I. synthesis of biosynthetic precursors and cofactors. *J. Theor. Biol.* **165,** 477–502 (1993).

107. Wiback, S. J., Famili, I., Greenberg, H. J. & Palsson, B. Ø. Monte Carlo sampling can be used to determine the size and shape of the steady-state flux space. *J. Theor. Biol.* **228,** 437–47 (2004).

108. Price, N. D., Schellenberger, J. & Palsson, B. O. Uniform sampling of steady-state flux spaces: means to design experiments and to interpret enzymopathies. *Biophys. J.* **87,** 2172–86 (2004).

109. Almaas, E., Kovács, B., Vicsek, T., Oltvai, Z. N. & Barabási, A.-L. Global organization of metabolic fluxes in the bacterium Escherichia coli. *Nature* **427,** 839–43 (2004).

110. Almaas, E., Oltvai, Z. N. & Barabasi, A. L. The activity reaction core and plasticity of metabolic networks. *PLoS Comput Biol* **1,** e68 (2005).

111. Samal, A., Matias Rodrigues, J. F., Jost, J., Martin, O. C. & Wagner, A. Genotype networks in metabolic reaction spaces. *BMC Syst. Biol.* **4,** 30 (2010).

112. Matias Rodrigues, J. F. & Wagner, A. Genotype networks, innovation, and robustness in sulfur metabolism. *BMC Syst Biol* **5,** 39 (2011).

113. Baba, T. *et al.* Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol. Syst. Biol.* **2,** 2006.0008 (2006).

114. Giaever, G. *et al.* Functional profiling of the Saccharomyces cerevisiae genome. *Nature* **418,** 387–91 (2002).

115. Papp, B., Pal, C. & Hurst, L. D. Metabolic network analysis of the causes and evolution of enzyme dispensability in yeast. *Nature* **429,** 661–664 (2004).

116. Vitkup, D., Kharchenko, P. & Wagner, A. Influence of metabolic network structure and function on enzyme evolution. *Genome Biol.* **7,** R39 (2006).

117. Bilu, Y., Shlomi, T., Barkai, N. & Ruppin, E. Conservation of expression and sequence of metabolic genes is reflected by activity across metabolic states. *PLoS Comput. Biol.* **2,** e106 (2006).

118. Pál, C., Papp, B. & Lercher, M. J. Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. *Nat. Genet.* **37,** 1372–5 (2005).

119. Moran, N. A. & Mira, A. The process of genome shrinkage in the obligate symbiont Buchnera aphidicola. *Genome Biol.* **2,** RESEARCH0054 (2001).

120. Pál, C. *et al.* Chance and necessity in the evolution of minimal metabolic networks. *Nature* **440,** 667–70 (2006).

121. Ibarra, R. U., Edwards, J. S. & Palsson, B. O. Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. *Nature* **420,** 186–9 (2002).

122. Fong, S. S., Marciniak, J. Y. & Palsson, B. Ø. O. Description and Interpretation of Adaptive Evolution of Escherichia coli K-12 MG1655 by Using a Genome-Scale In Silico Metabolic Model. *J. Bacteriol.* **185,** 6400–6408 (2003).

123. Fong, S. S. & Palsson, B. Ø. Metabolic gene-deletion strains of Escherichia coli evolve to computationally predicted growth phenotypes. *Nat. Genet.* **36,** 1056–8 (2004).

124. Segrè, D., Vitkup, D. & Church, G. M. Analysis of optimality in natural and perturbed metabolic networks. *Proc. Natl. Acad. Sci. U. S. A.* **99,** 15112–7 (2002).

125. Shlomi, T., Berkman, O. & Ruppin, E. Regulatory on/off minimization of metabolic flux changes after genetic perturbations. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 7695–700 (2005).

126. Park, J. H., Lee, K. H., Kim, T. Y. & Lee, S. Y. Metabolic engineering of Escherichia coli for the production of L-valine based on transcriptome analysis and in silico gene knockout simulation. *Proc. Natl. Acad. Sci. U. S. A.* **104,** 7797–802 (2007).

127. Lee, K. H., Park, J. H., Kim, T. Y., Kim, H. U. & Lee, S. Y. Systems metabolic engineering of Escherichia coli for L-threonine production. *Mol. Syst. Biol.* **3,** 149 (2007).

128. Moon, S. Y., Hong, S. H., Kim, T. Y. & Lee, S. Y. Metabolic engineering of Escherichia coli for the production of malic acid. *Biochem. Eng. J.* **40,** 312–320 (2008).

129. Fong, S. S. *et al.* In silico design and adaptive evolution of Escherichia coli for production of lactic acid. *Biotechnol. Bioeng.* **91,** 643–8 (2005).

130. Wang, Q., Chen, X., Yang, Y. & Zhao, X. Genome-scale in silico aided metabolic analysis and flux comparisons of Escherichia coli to improve succinate production. *Appl. Microbiol. Biotechnol.* **73,** 887–94 (2006).

131. Lee, S. J. *et al.* Metabolic engineering of Escherichia coli for enhanced production of succinic acid, based on genome comparison and in silico gene knockout simulation. *Appl. Environ. Microbiol.* **71,** 7880–7 (2005).

132. Alper, H., Jin, Y.-S., Moxley, J. F. & Stephanopoulos, G. Identifying gene targets for the metabolic engineering of lycopene biosynthesis in Escherichia coli. *Metab. Eng.* **7,** 155–64 (2005).

133. Alper, H., Miyaoku, K. & Stephanopoulos, G. Construction of lycopene-overproducing E. coli strains by combining systematic and combinatorial gene knockout targets. *Nat. Biotechnol.* **23,** 612–6 (2005).

134. Haverkorn van Rijsewijk, B. R. B., Nanchen, A., Nallet, S., Kleijn, R. J. & Sauer, U. Large-scale 13C-flux analysis reveals distinct transcriptional control of respiratory and fermentative metabolism in Escherichia coli. *Mol. Syst. Biol.* **7,** 477 (2011).

135. El-Mansi, M., Cozzone, A. J., Shiloach, J. & Eikmanns, B. J. Control of carbon flux through enzymes of central and intermediary metabolism during growth of Escherichia coli on acetate. *Curr. Opin. Microbiol.* **9,** 173–9 (2006).

136. Byrne, D., Dumitriu, A. & Segrè, D. Comparative multi-goal tradeoffs in systems engineering of microbial metabolism. *BMC Syst. Biol.* **6,** 127 (2012).

137. Burgard, A. P., Pharkya, P. & Maranas, C. D. Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnol. Bioeng.* **84,** 647–57 (2003).

138. Yim, H. *et al.* Metabolic engineering of Escherichia coli for direct production of 1,4-butanediol. *Nat. Chem. Biol.* **7,** 445–52 (2011).

139. Pharkya, P., Burgard, A. P. & Maranas, C. D. OptStrain: a computational framework for redesign of microbial production systems. *Genome Res.* **14,** 2367–76 (2004).

140. Rocha, I. *et al.* OptFlux: an open-source software platform for in silico metabolic engineering. *BMC Syst. Biol.* **4,** 45 (2010).

141. Ranganathan, S., Suthers, P. F. & Maranas, C. D. OptForce: an optimization procedure for identifying all genetic manipulations leading to targeted overproductions. *PLoS Comput. Biol.* **6,** e1000744 (2010).

142. Tepper, N. & Shlomi, T. Predicting metabolic engineering knockout strategies for chemical production: accounting for competing pathways. *Bioinformatics* **26,** 536–43 (2010).

143. Pharkya, P. & Maranas, C. D. An optimization framework for identifying reaction activation/inhibition or elimination candidates for overproduction in microbial systems. *Metab. Eng.* **8,** 1–13 (2006).

144. Patil, K. R., Rocha, I., Förster, J. & Nielsen, J. Evolutionary programming as a platform for in silico metabolic engineering. *BMC Bioinformatics* **6,** 308 (2005).

145. Raman, K., Rajagopalan, P. & Chandra, N. Flux balance analysis of mycolic acid pathway: targets for anti-tubercular drugs. *PLoS Comput. Biol.* **1,** e46 (2005).

146. Colijn, C. *et al.* Interpreting expression data with metabolic flux models: predicting Mycobacterium tuberculosis mycolic acid production. *PLoS Comput. Biol.* **5,** e1000489 (2009).

147. Jamshidi, N. & Palsson, B. Ø. Investigating the metabolic capabilities of Mycobacterium tuberculosis H37Rv using the in silico strain iNJ661 and proposing alternative drug targets. *BMC Syst. Biol.* **1,** 26 (2007).

148. Kim, T. Y., Kim, H. U. & Lee, S. Y. Metabolite-centric approaches for the discovery of antibacterials using genome-scale metabolic networks. *Metab. Eng.* **12,** 105–11 (2010).

149. Heinemann, M., Kümmel, A., Ruinatscha, R. & Panke, S. In silico genome-scale reconstruction and validation of the Staphylococcus aureus metabolic network. *Biotechnol. Bioeng.* **92,** 850–64 (2005).

150. Lee, D.-S. *et al.* Comparative genome-scale metabolic reconstruction and flux balance analysis of multiple Staphylococcus aureus genomes identify novel antimicrobial drug targets. *J. Bacteriol.* **191,** 4015–24 (2009).

151. Raghunathan, A. *et al.* In Silico Metabolic Model and Protein Expression of Haemophilus influenzae Strain Rd KW20 in Rich Medium. *OMICS* **8,** 25–41 (2004).

152. Oberhardt, M. A., Puchałka, J., Fryer, K. E., Martins dos Santos, V. A. P. & Papin, J. A. Genome-scale metabolic network analysis of the opportunistic pathogen Pseudomonas aeruginosa PAO1. *J. Bacteriol.* **190,** 2790–803 (2008).

153. Schilling, C. H. *et al.* Genome-scale metabolic model of Helicobacter pylori 26695. *J. Bacteriol.* **184,** 4582–93 (2002).

154. AbuOun, M. *et al.* Genome scale reconstruction of a Salmonella metabolic model: comparison of similarity and differences with a commensal Escherichia coli strain. *J. Biol. Chem.* **284,** 29480–8 (2009).

155. Baart, G. J. E. *et al.* Modeling Neisseria meningitidis metabolism: from genome to metabolic fluxes. *Genome Biol.* **8,** R136 (2007).

156. Suthers, P. F. *et al.* A genome-scale metabolic reconstruction of Mycoplasma genitalium, iPS189. *PLoS Comput. Biol.* **5,** e1000285 (2009).

2. Superessential reactions in metabolic networks

Aditya Barve[1,3], João F. Matias Rodrigues[2,3] and Andreas Wagner [1,3,4]

[1] Institute of Evolutionary Biology and Environmental Sciences, Bldg. Y27, University of Zurich, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland

[2] Institute of Molecular Life Sciences, University of Zurich, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland

[3] The Swiss Institute of Bioinformatics, Bioinformatics, Quartier Sorge, Batiment Genopode, 1015 Lausanne, Switzerland.

[4] The Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA

Part of this chapter was published in *Proceedings of National Academy of Sciences. USA* 109, E1121–E1130 (2012) [doi: 10.1073/pnas.1113065109]

2.1 Abstract

The metabolic genotype of an organism can change through loss and acquisition of enzyme-coding genes, while preserving its ability to survive and synthesize biomass in specific environments. This evolutionary plasticity allows pathogens to evolve resistance to antimetabolic drugs, by acquiring new metabolic pathways that by-pass an enzyme blocked by a drug. We here study quantitatively the extent to which individual metabolic reactions and enzymes can be by-passed. To this end, we use a recently developed computational approach to create large metabolic network ensembles that can synthesize all biomass components in a given environment, but that contain an otherwise random set of known biochemical reactions. Using this approach, we identify a small connected core of 124 reactions that are absolutely superessential, that is, required in all metabolic networks. Many of these reactions have been experimentally confirmed as essential in different organisms. We also report a superessentiality index for thousands of reactions. This index indicates how easily a reaction can be by-passed. We find that it correlates with the number of sequenced genomes that encode an enzyme for the reaction. Superessentiality can help choose an enzyme as a potential drug target, especially since the index is not highly sensitive to the chemical environment a pathogen requires. Our work also shows how analyses of large network ensembles can help understand the evolution of complex and robust metabolic networks.

2.2 Introduction

The metabolic networks of free-living organisms are complex and comprise hundreds to thousands of chemical reactions. Most of these reactions are catalyzed by enzymes encoded in genes. A metabolic network's most important function is to synthesize all small molecule precursors of biomass that are necessary for the growth and survival of an organism. For well-studied free-living organisms these comprise some 50 different small molecules, including amino acids and nucleotides [1].

The metabolic genotype of an organism comprises all enzyme-coding genes. It determines the enzymatic reactions in a metabolic network. This genotype can change dramatically without affecting the metabolic phenotype, that is, the ability to synthesize biomass in a given environment. For instance, loss-of-function mutations in many enzyme-coding genes can leave the metabolic phenotype unaffected [2–8]. In addition, reactions can get added to a metabolic network through horizontal gene transfer of enzyme-coding genes, a process that is especially frequent in prokaryotes. The deletion and addition of multiple reactions over time may lead to metabolic networks that differ in many reactions, but that can still sustain life in the same chemical environment.

This enormous genotypic plasticity has implications for the evolution of metabolism. It means that reactions or entire pathways necessary for life in one organism may be dispensable in another organism. For example, the isoprenoid pathway synthesizes isopentenyl diphosphate, which is important for synthesis of cell wall constituents. This pathway is essential in *Bacillus subtilis*, but it is replaced by the mevalonate pathway in *Staphylococcus aureus,* where the mevalonate pathway is essential [9,10]. Neither of the pathways would be essential in an organism possessing both of these metabolic routes. For the purpose of our work, we define a biochemical reaction as *essential* if its elimination abolishes the network's ability to synthesize all biomass molecules in a given environment. A reaction is non-essential if an organism has the ability to bypass that reaction through alternate reactions or metabolic pathways, or if the product of the reaction is not needed in a given environment. We emphasize that

reaction essentiality depends on the environment. Earlier analyses [11–15] have explored to which extent reaction essentiality varies among environments. However, these studied focused on a single metabolic network and its genotype. They did not take into account that metabolic networks with the same phenotype can vary in their genotype. Such genotypic variation can also lead to variation in reaction and gene essentiality. The reactions in the isoprenoid and mevalonate pathways mentioned above provide one example. Another is the existence of essential genes unique to particular strains of *S. cerevisiae* [16].

Enormous investments are necessary to develop new antibiotic drugs that combat pathogens [17].The genotypic plasticity of metabolic networks has very practical implications for these efforts, and for the long-term success of the drugs they produce. Multiple existing drugs target the metabolism of pathogens, such as sulfonamides, fosmidomycin and isoniazid [18–21]. An ideal enzymatic drug target has to fulfill several criteria, among them that the target is essential for the pathogen's survival. Only in this case can the drug suppress the pathogen. However, as we pointed out, whether an enzyme is essential may depend on the metabolic network it is a part of. The same enzyme can be essential in one metabolic network that can sustain life in a given environment, but non-essential in a different network. Drugs targeting such enzymes are vulnerable to pathogens that evolve resistance against them, for example through horizontal gene transfer.

Which reactions in a metabolic network may be most easily by-passed? Which reactions cannot be by-passed? We do not know the answer to these questions. This is not surprising. Answering them would require examining many different metabolic genotypes, and evaluating the essentiality of reactions in each of them. This cannot be done systematically with current experimental technology, and requires new computational approaches. We have recently developed an approach that can answer these questions [22,23]. It uses flux balance analysis (FBA) to compute the *phenotype* of a network from its genotype. This phenotype is the ability of the network to sustain life in an environment or a set of environments. FBA has been shown to predict gene essentiality with an accuracy of nearly 90 percent [12,13]. Mismatches between FBA

predictions and experimental data can often be attributed to enzyme misregulation, wherein regulatory constraints prevent enzymes to be expressed at optimal levels [24–26]. Such constraints are easily broken in laboratory evolution experiments [24–26] and are of limited relevance to our work, because we are concerned with a more fundamental question, namely how the presence or absence of some reactions (enzyme-coding genes) affects the essentiality of other reactions.

Even more central to our approach than FBA is our current considerable knowledge of the "universe" of biochemical reactions. This known universe currently comprises more than 5,000 stoichiometrically defined reactions [27,28] that are known to take place in some organisms. Based on this information, our approach can generate random samples of metabolic genotypes (metabolic networks) with a given phenotype [22,23] (see methods, section 2.5). We refer to such genotypes as *random viable* metabolic networks. Briefly, we here generate large samples of such networks, examine the reactions in them, and to determine whether they are essential. We then use the concept of reaction superessentiality [23] to estimate a superessentiality index for each reaction. This index indicates the fraction of random metabolic networks with a given phenotype in which a reaction is essential. Reactions where this index is low are easily by-passed, reactions where this index is high are difficult to by-pass, and reactions where this index is maximal are impossible to by-pass based on our current knowledge. The word superessentiality is motivated by the fact that reactions can be more than just essential. They can be essential in many, most, or all metabolic networks with a given phenotype. Our analysis focuses on carbon metabolism, because carbon is central to life.

In this context, we ask fundamental questions about essential reactions and the extent to which they are also superessential. Which are the reactions that cannot be by-passed? How many are they? To what extent does their essentiality depend on the environment? We relate the outcome of these and other analyses to metabolic evolution and to the problem of finding drug targets with high-superessentiality and thus low propensity for resistance evolution.

2.3 Results

A core of absolutely superessential reactions in carbon metabolism

We begin our analysis with a single carbon source phenotype, an aerobic minimal environment that contains glucose as the *only* carbon source (see section 2.5 for all environmental metabolites we study). Our point of departure is the biomass composition of *E. coli*, because it is well-understood and its major components are representative of other free-living organisms [1,12]. Starting from the set of all currently known reactions, we generated a metabolic network that we call the universal network. This network comprises all known 5906 biochemical reactions with well-defined stoichiometry (see , section 2.5 for construction of the universal network). Not surprisingly, this network can produce all biomass components in a glucose minimal environment. For any one metabolic reaction, the universal network contains all possible alternative pathways that could by-pass a reaction. Because the reactions of any viable network (including the *E. coli* network) are a subset of this universe of reactions, an essential reaction in this network cannot be by-passed in *any* network that uses reactions from the known universe of reactions. That is, if a reaction is essential in the universal network, no known pathway can by-pass this reaction and render it non-essential. We analyzed each reaction in the universal network for its essentiality, and thus identified 133 reactions essential for growth on glucose. This set of 133 essential reactions forms an irreducibly essential set of reactions. We call it the "superessential core" of metabolism for viability on glucose (see supplementary material S1 in [29]).

A broad distribution of reaction superessentiality

As opposed to reactions in the superessential core, which are essential regardless of which other reactions occur in a network, the essentiality of many reactions may depend on other reactions. Although the universal network allowed us to identify absolutely superessential reactions, it does not allow us to understand how reaction essentiality depends on other reactions in a network. To this end, we took a different approach, and evaluated the essentiality of each reaction in a large number of *genome-scale* metabolic networks that contain an otherwise random assortment of known

reactions, but are viable on a given set of environments. Starting from the *E. coli* metabolic network, we used the approach detailed in methods (section 2.5) to generate random samples of metabolic networks that can synthesize all *E. coli* biomass precursors in an aerobic minimal environment containing glucose as the only carbon source. Briefly, this approach relies on Markov Chain Monte Carlo sampling from the set of all metabolic networks that can be formed using 5609 known biochemical reactions. Our method ensures that the resulting networks have the same number of reactions, but are otherwise unrelated to the starting network, and have a randomized reaction content relative to each other. We refer to these networks as *random viable* networks. We generated 500 such random viable networks, and identified all essential reactions in each such network. These are reactions whose removal abolishes a network's ability to synthesize all biomass components in this environment. On average, only 283.59 reactions (20.3 percent) were essential in networks of our sample, with a low standard deviation of 8.51 reactions.

To quantify a reaction's superessentiality with this approach, we define its *superessentiality index* (I_{SE}) as the fraction of networks in which the reaction is essential. The maximum value of I_{SE} is one for reactions that are essential in all networks of the sample – we call such reactions *absolutely superessential*. The lowest value of I_{SE} is zero for reactions nonessential in all networks. An I_{SE} of "0.002" would indicate that a reaction was essential only in one out of 500 random viable networks. Out of the total number of 5906 chemical reactions that occurred in at least one of our 500 random viable networks (see methods, section 2.5), 1400 (23.7 percent) reactions were essential in at least one network. Figure 2.1 shows a rank plot in which reactions are ranked according to their superessentiality index. It indicates that different reactions can vary widely in their superessentiality.

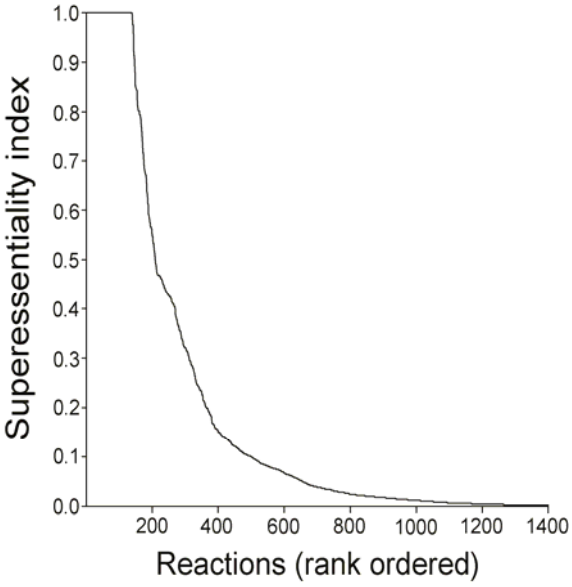


Figure 2.1: Rank plot of the superessentiality index of 1400 reactions essential for growth on glucose in 500 random viable metabolic networks. The plateau to the left of the plot corresponds to 139 absolutely superessential reactions ($I_{SE} = 1$). Data is based on essential reactions from 500 random metabolic networks viable on glucose.

Comparing the results of this approach to our previous determination of the superessential core from the universal network allows us to validate the network sampling approach. Specifically, sampling identified 139 (2.3 percent) of reactions as absolutely superessential ($I_{SE} = 1$), that is, they cannot be by-passed in any of our 500 random networks viable on glucose. These reactions correspond to the plateau on the left-hand side of Figure 2.1. Based on the universal network, we had found that 133 reactions formed the superessential core for viability on glucose. Most importantly, these 133 reactions are all contained in the set of 139 absolutely superessential reactions identified by sampling (supplementary material S1 in [29]). In other words, only six reactions that sampling identified as absolutely superessential are artifactually identified as absolutely superessential because of insufficient sampling. This observation shows that even modest samples of 500 random metabolic networks can provide good estimates of reaction superessentiality.

How many of the essential reactions in *E. coli* can potentially be by-passed by reactions from the reaction universe? 240 reactions are essential in *E.coli* for growth in the glucose minimal environment, of which 133 are absolutely superessential in the universal network. This means that 55.4 percent (133 of 240) of the essential reactions from *E.coli* are in fact absolutely superessential and thus irreplaceable. The remaining

44.6 percent reactions have an $I_{SE} < 1$, meaning that an organism could bypass such reactions by acquiring new metabolic genes through mechanisms such as horizontal gene transfer.

Examples of superessential reactions

We next discuss a few examples of reactions in the superessential core. One of them is phosphoglucosamine mutase (*glmM*; Blattner number b3176) [30], which catalyzes a reversible conversion between glucosamine-1-phosphate to glucosamine-6-phoshate. This enzyme plays an important role in the synthesis of UDP-*N*-acetyl-D-glucosamine, which is used in peptidoglycan and lipid IV_A biosynthesis [31]. Second, nicotinamide adenine dinucleotide (NAD) kinase (*nadK*; Blattner number b2615) is important in the generation of nicotinamide adenine dinucleotide phosphate (NADP) from NAD in an ATP dependent manner. NAD kinase thus may play an important role in determining the size of a cell's NADP pool and its turnover in the cell [32]. A third example is diaminopimelate decarboxylase (*lysA*; Blattner number b2838), which generates L-lysine from *meso*-diaminopimelate. This enzyme catalyzes the last reaction in the L-lysine biosynthesis pathway. It is essential if L-lysine is not supplied by the environment.

In addition to absolutely superessential reactions ($I_{SE} = 1$), reactions with lower superessentiality index ($I_{SE} < 1$) can also shed light on the structure of metabolism. For example, if a reaction is non-essential in a fraction ($1 - I_{SE}$) of random viable networks, this fraction indicates how easily the reaction can be by-passed by alternate metabolic pathway(s) based on known reactions. For instance, glucose-6-phosphate isomerase (*pgi*; Blattner number b4025) while not essential in *E.coli* [11], has an I_{SE} of 0.314, indicating that it is essential in 31.4 percent of networks. This means that it is by-passed in 68.6 percent of the networks in our sample. Our analysis of reactions shows that reactions from central pathways such as glycolysis, citric acid cycle or pyruvate metabolism tend to have low superessentiality indices, while reactions from amino acid synthesis, such as histidine metabolism tend to have especially high superessentiality indices (see supplementary material S2 in [29] and section 2.6).

Individual networks may contain reactions that do not contribute to biomass production, for example, because they are part of an isolated pathway fragment, or a pathway that does not contribute to biomass synthesis in a given environment. Such reactions and pathway fragments do occur in well-annotated metabolic models like that of *E. coli* [12]. They are also inevitable consequences of unbiased Markov Chain Monte Carlo sampling of random viable networks (see methods, section 2.5). We refer to such reactions as blocked reactions [23,33]. To identify them, we computed for each network in our sample of 500 random viable networks the maximum allowable flux through each reaction for viability on glucose (see methods, section 2.5) [33]. If this flux was below a threshold (10^{-8}), we consider the reaction "blocked". When considered together, the number of networks in which a reaction occurs, its superessentiality index I_{SE}, and the number of networks it is blocked in can indicate the extent to which a reaction and its alternate pathways coexist and are functional in a sample of random networks. As mentioned earlier, glucose-6-phosphate isomerase is essential in 31.4 percent of networks. However, it is present in 44.2 percent of networks (and blocked in none). Together these proportions mean that 12.8 percent (44.2 – 31.4 – 0) of the random networks in our sample have more than one functional route(s) for this particular reaction. The penultimate reaction in histidine biosynthesis is carried out by histidinol dehydrogenase (*hisD*, Blattner number b2020). It is present in 87 percent of the networks, essential in 84.6 percent, and blocked in 0.6 percent of networks, meaning that it coexists along with its alternate pathway(s) only in 1.8 percent (87 – 84.6 – 0.6) of networks. This measure of superessentiality is complementary to the I_{SE} index in providing information on how easily a reaction is by-passed. We report it for all *E. coli* reactions in supplementary material S2 in [29].

Metabolic networks have many environment-general essential reactions

Thus far, we discussed reaction essentiality for a single carbon source phenotype. How do our observations generalize to multiple carbon source phenotypes? Our definition of phenotype regards the ability of a network to sustain life on a given number of *sole* carbon sources. Networks with a multiple carbon source phenotype can sustain life on many sole carbon sources, and on any subset of these carbon sources, such as might occur in an environment that changes over time. The highest number of carbon sources for a multiple carbon source phenotype we consider are the

54 different sole carbon sources in which *E. coli* is known to be viable from experiments (see methods, section 2.5) [12] . We can represent the phenotype of viability of 54 carbon sources as a binary string of length 54 in which all entries are equal to one. A deletion of a reaction that abolishes viability on carbon source *i* would change the value of entry *i* in this string to zero. We define a reaction as essential in this multiple carbon source phenotype if it abolishes viability on at least one carbon source. Deletion of some reactions abolishes viability in all 54 environments. We refer to such reactions as *environment-general* essential reactions. Deletion of other reactions abolishes viability only in a few environments. We refer to these reactions as *environment-specific* essential reactions.

We next revisit for a multiple carbon phenotype the concept of a superessential core of metabolism, the set of absolutely superessential reactions. As in our analysis with the universal network for a single carbon source phenotype, we identified absolutely superessential reactions for growth on all 54 carbon environments from our universal reaction network of 5906 reactions. This approach yielded 148 absolutely superessential reactions. We note that only 15 additional reactions became absolutely superessential as our analysis moved from the single carbon source to the multiple carbon source phenotype (148 versus 133 absolutely superessential reactions). This observation argues for a common core of superessential reactions that does not depend on the actual environment considered. Indeed, we find that 125 of the 148 reactions in the superessential core are environment-general, meaning that deleting these reactions abolishes growth in all 54 different environments.

We next also identified all essential reactions (as defined above) with our sampling approach, that is, in each of 500 random networks that were viable on 54 carbon sources, and determined each reaction's superessentiality index (I_{SE}) (We note that this index disregards the number of environments in which a reaction is essential). The number (1569) of reactions that were essential in at least one network was only 12 percent higher than the 1400 reactions essential for growth on glucose that we discussed earlier. Figure S1 (section 2.6) shows a rank plot of I_{SE} for these 1569 reactions. Its shape is very similar to the curve in Figure 2.1, and its left-hand side also contains a plateau of 155 (9.9 percent) reactions that were absolutely

superessential for growth on each of the 54 carbon sources. In sum, only seven more reactions were identified as absolutely superessential through the sampling approach in comparison to the 148 reactions from the universal network. The 148 true-positive reactions are shown in supplementary material 3 in [29].

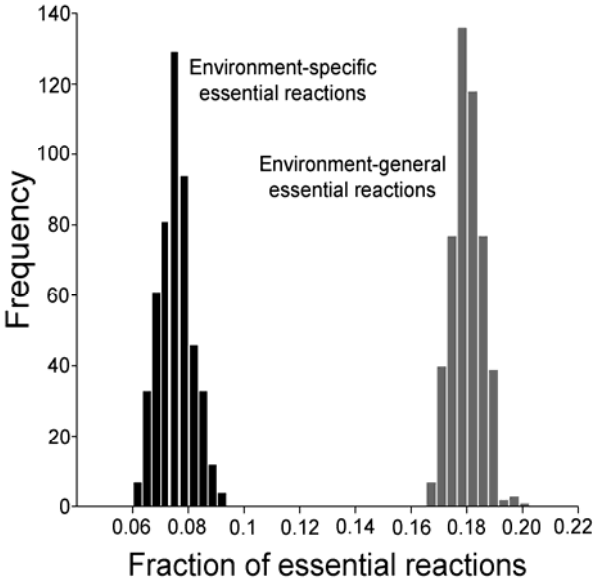
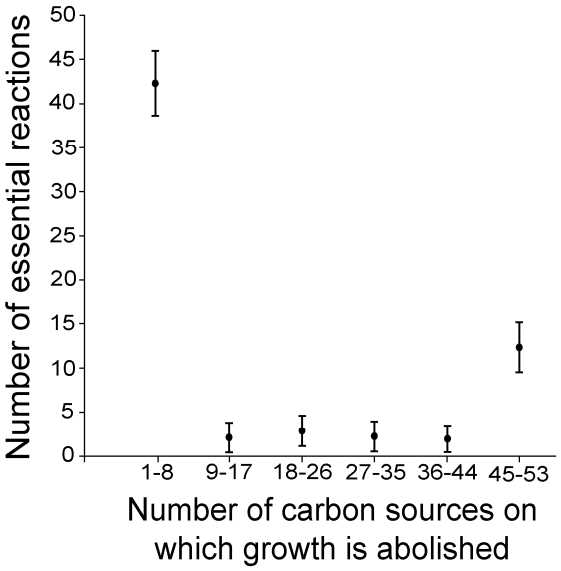


Figure 2.2: The distribution of environment-specific and environment-general essential reactions in 500 random metabolic networks viable on 54 carbon sources. The mean proportion of environment-general essential reactions for all 500 networks is 2.38 times higher than the mean proportion of environment-specific essential reactions. Data is based on essential reactions from 500 random metabolic networks viable on 54 carbon sources.

Figure 2.3: The vertical axis shows the mean number of reactions whose elimination abolishes viability on the number of carbon sources shown on the horizontal axis. Error bars correspond to one standard deviation. Deletion of most essential reactions abolishes growth only on a very limited number of carbon sources (1-8). Data is based on 500 random networks viable on 54 carbon sources.



While deletion of some reactions abolishes growth on one or few environments, deletion of other reactions abolishes growth on all 54 environments. Which of these two types of essential reaction is more prominent? Figure 2.2 shows the distribution

of the proportion of a metabolic network's reactions that are environment-specific and environment-general. The data is based on 500 random networks viable on 54 different sole carbon sources. The figure clearly shows that many more essential reactions are environment-general (mean = 0.18 or 251 reactions) than environment-specific (mean = 0.0755 or 105 reactions). That is, most essential reactions abolish viability on all carbon sources.

Figure 2.3 shows the average number of environment-specific essential reactions per network whose deletion abolishes viability on a given range of carbon sources. The number of reactions whose deletion abolishes growth on fewer than 8 sources is by far the highest, about 42 per network (s.e.m.: 4 reactions), while reactions that abolish growth on 45 to 53 sources are next, with 12 per network (s.e.m.: 3 reactions). Thirdly, there are fewer reactions (1-3 per network, s.e.m.: 1 reaction) that abolish growth on 9-45 environments. In sum, most essential reactions in a metabolic network fall into two categories, those whose deletion abolishes viability in very few environments, and the vast majority whose deletion abolishes viability in all environments.

The superessentiality index of a reaction is not very sensitive to the environment

Our analysis thus far also suggests that most essential reactions are essential irrespective of the number of different environments and regardless of the specific carbon source examined. For example, supplementary figure S1 (see section 2.6) shows that 1569 reactions are essential in at least one network for viability on at least one of 54 alternative carbon sources, while figure 2.1 shows that 1400 reactions are essential in at least one network for growth on glucose. Thus, out of 1655 unique reactions that are essential for life on either glucose or on at least one of 54 carbon sources, 1314 reactions (79.4 percent) are essential for both kinds of phenotypes. In addition, the superessentiality of reactions is similar for both the single carbon source and the multiple carbon source phenotype. Figure 2.4 shows that a strong correlation (Pearson's r = 0.95, p-value < 10^{-300}, n = 1314) exists between the superessentiality index I_{SE} in the single and the multiple carbon source phenotypes. A comparison between essential reactions of the phenotype requiring growth on 54 carbon sources,

and simpler phenotypes that require growth on 5, 10, 20, 30, and 40 alternative carbon sources yields correlations as high as that seen in figure 2.4 (Pearson's r > 0.92 , *p*-value < 10^{-300}, n ≥ 1409 in all five cases). Moreover, as we already discussed, most essential reactions in a network are essential for growth in more than one environment (figures 2.3 and 2.4). Taken together, all this means that a reaction's superessentiality index I_{SE} does not depend strongly on the number of carbon sources on which it can support viability. The same holds therefore for how readily a reaction can be by-passed: it does not depend strongly on the environment for most reactions. In the supplementary results (section 2.6), we discuss some notable exceptions, reactions that can have high superessentiality in 54 different minimal environments but low superessentiality in glucose minimal environment.

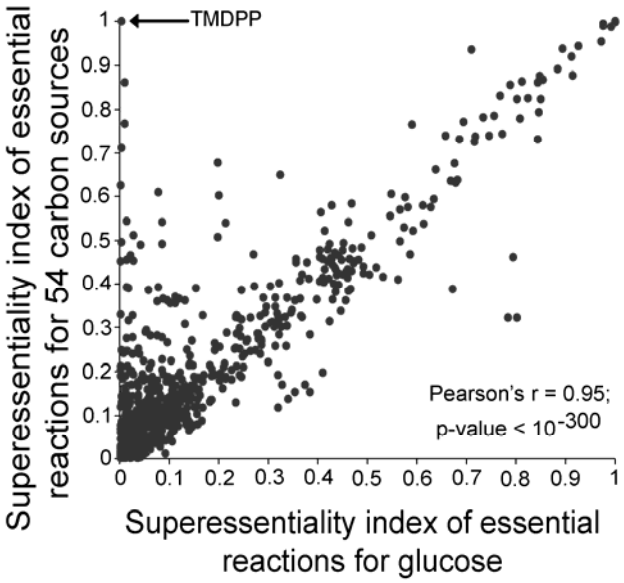


Figure 2.4: Superessentiality indices of reactions essential for growth on glucose (horizontal axis) and on 54 carbon sources (vertical axis). The strong association demonstrates that superessentiality does not depend strongly on the environment. One reaction whose superessentiality differs dramatically in glucose and 54 carbon sources is indicated with an arrow (TMDPP, discussed in section 2.6). Data is based on essential reactions from 500 random metabolic networks viable on glucose, and on another set of 500 random metabolic networks viable on 54 carbon sources.

Absolutely superessential reactions are enriched in anabolic pathways

As mentioned earlier, we found 133 reactions that are absolutely superessential for growth on glucose and 125 reactions that are absolutely superessential and environment-general for growth on 54 alternative sole carbon sources. We asked whether reactions in these two superessential cores preferentially derive from specific pathways (see methods, section 2.5). We found that both cores were significantly enriched for reactions in pathways that synthesize several amino acids (histidine, valine, leucine, isoleucine, tyrosine, and tryptophan) and cell wall components (see

methods, section 2.5). In contrast, the cores were not enriched for pathways that synthesize murein, threonine, lysine, methionine, as well as membrane lipids. The results are similar for both superessential cores (supplementary tables S2 and S3, section 2.6).This analysis also showed that reactions from central metabolism such as glycolysis or the citric acid cycle are notably absent from the superessential cores (supplementary materials S1 and S3 in [29]). Taken together, this means that most absolutely superessential reactions are anabolic in nature, whereas catabolic reactions from pathways such as glycolysis can be more easily by-passed. This observation agrees with experimental and computational studies of essential reactions in *E. coli* and *S. cerevisiae* [6,11,34,35]. It is also consistent with our earlier observation that reactions from these pathways generally have low superessentiality indices (supplementary material S2 in [29]). We speculate that the reticulate structure of some parts of metabolism may be the reason why one or more pathways, such as central carbon metabolism, are not enriched for superessential reactions, even though these pathways are very important ([36] and supplementary materials S1 and S3 in [29]). In contrast, some amino acids such as histidine or, tryptophan are synthesized through more linear pathways which may thus not be as easy to bypass. In section 2.6, we show that the environment-general superessential core is a compact and highly connected part of metabolism.

Genes responsible for superessential reactions occur in most prokaryotic genomes

If a reaction is frequently essential for preserving a phenotype in random viable metabolic networks, then the corresponding enzyme-coding gene should also occur frequently in many prokaryotic genomes. This line of reasoning will fail if either our knowledge of the universe of metabolic reactions is incomplete, or if our understanding of an organism's enzyme complement is partial. The extent to which it fails can shed light on how imperfect our current metabolic knowledge is. With these observations in mind, we analyzed the relationship between reaction superessentiality and reaction occurrence in prokaryotic genomes. To this end, we defined the *genome-occurrence index* (I_{GO}) of a reaction as the fraction of genomes that carry a gene whose product is known to catalyze this reaction. For each reaction in the universe of reactions, we used Kegg Orthology (KO) numbers

(http://www.genome.jp/kegg/ko.html) [37] to estimate the fraction of prokaryotic genomes that encode an enzyme catalyzing the reaction (see methods, section 2.5).

We first focused on absolutely superessential reactions ($I_{SE} = 1$) for a single carbon source phenotype (viability on glucose) and for multiple carbon source phenotypes. Figure 2.5 shows a rank plot of the genome occurrence index I_{GO} for these reactions. If our current understanding of metabolism was perfect, we would expect that the I_{GO} values of all absolutely superessential genes are equal to one. But the rank plot shows that this is not the case. We highlight two features of this plot. First, the plots for single carbon source and multiple carbon source phenotypes are very similar and nearly congruent. This corroborates our earlier observation that most superessential reactions are environment-general.

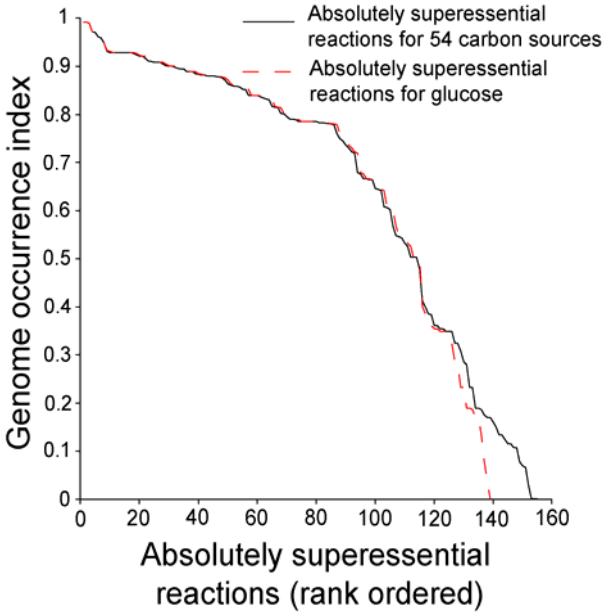


Figure 2.5: The vertical axis shows the genome occurrence index of absolutely superessential reactions for growth on glucose and 54 carbon sources. The curves are almost congruent, underscoring that most absolutely superessential reactions are environment-general. Furthermore, a majority (73.5 percent) of the 155 absolutely superessential reactions for growth on 54 carbon sources have a high genome occurrence ($I_{GO} \geq 0.5$).

Figure 2.5 suggests that differences in the environment in which different species live cannot account for most differences in genome occurrences among absolutely superessential reactions. Second, out of the 155 absolutely superessential reactions for growth in 54 carbon sources, 57 percent (88 reactions) occur in more than 75 percent of genomes (820 genomes); 73.5 percent (114) of these 155 reactions occur in more

than 50 percent of genomes ($I_{GO} \geq 0.5$). This means that a majority of absolutely superessential reactions occur in most prokaryotic genomes sequenced to date. This association between superessentiality and genome occurrence (I_{GO}) is highly significant with a *p*-value smaller than 10^{-5} ($n = 10^5$, permutation test) (supplementary figure S3 in section 2.6). The 125 absolutely superessential and environment-general reactions also showed a highly significant association with genome occurrence. Specifically, out of 125 reactions, 105 reactions (84 percent) occurred in more than 50 percent of genomes (*p*-value $< 10^{-5}$, $n = 10^5$). We next expanded our analysis to include reactions with lower than absolute superessentiality ($I_{SE} < 1$), and determined whether there exists a statistical association between a reaction's superessentiality index and its genome occurrence. Such an association indeed exists (Spearman's $\rho = 0.4$, *p*-value $< 10^{-300}$, $n = 5609$). The supplementary figure S4 (section 2.6) shows that this correlation in the observed data is significantly different than that in randomized data (*p*-value $< 10^{-5}$, $n = 10^5$), suggesting that the association we see between superessentiality index and the occurrence of a reaction's enzyme-coding genes does not occur by chance alone.

These observations indicate, on the one hand, that our approach of characterizing reaction superessentiality in random viable networks reveals biologically relevant information. On the other hand, they also show that our knowledge of metabolism and its enzymes is still incomplete. For example, as we discussed earlier, 114 of 155 absolutely superessential reactions occur in at least 50 percent of genomes. Of the remaining 41 reactions, about one half are environment-specific reactions, while other reactions have low genome occurrences. The reasons for apparently low occurrence of highly superessential reactions highlight various limitations in existing genome annotation and database information, as a few examples will show.

Glycerol-3-phosphate acyltransferase, encoded by *plsB*, plays a role in phospholipid synthesis and is essential in *E. coli* [11] and *S. typhirium* [38] in the murine *in-vivo* infection model. The enzyme utilizes fatty acyl-ACP or acyl-CoA thioesters to form acyl-glycerol-3-phosphate, and is mainly limited to Gamma-proteobacteria [39], which explains its low genome occurrence index. Other prokaryotes contain a similar

enzyme that uses acyl-phosphate to synthesize acyl-glycerol-3-phosphate (encoded by the gene *plsY*), but the KEGG reaction database does not distinguish between these reactions. It only contains the *E. coli* variant. If both variants were taken into account, the reaction would occur in a much larger fraction (0.91) of genomes. The glycerol-3-phosphate acyltransferase reaction appears superessential because the other pathway (encoded by *plsY*) is absent from the set of known reactions we used. Another example concerns the *coaA* gene, encoding the enzyme pantothenate kinase. Pantothenate kinase catalyzes the first step of Coenzyme A biosynthesis, an essential and ubiquitous cofactor in almost all biological organisms. Prokaryotes can have two different types of pantothenate kinases, encoded by the gene *coaA* or *coaX*. These genes do not share sequence similarity [40], but they nonetheless encode enzymes that catalyze the same reaction. The enzyme encoded by *coaA* is of type I, while the enzyme encoded by *coaX* is a type III enzyme [40,41]. The existence of the alternative *coaX* explains the low genome occurrence of *coaA*, but it also reinforces the essentiality of the reaction. If we take both variants into account, the reaction occurs in 83 percent of the genomes we analyzed. Another case in point is reactions catalyzed by promiscuous enzymes, such as the enzyme pyrimidine phosphatase (PMDPHT). PMDPHT catalyzes the transformation of 5-amino-6-(5'-phosphoribitylamino) uracil to 4-(1-D-Ribitylamino)-5-aminouracil with the release of one inorganic phosphate [12] and is essential in the riboflavin biosynthesis pathway [1]. PMDPHT is absolutely superessential and environment-general, but the enzyme-coding gene responsible for PMDPHT has not yet been identified. It therefore has the minimum genome occurrence of zero.

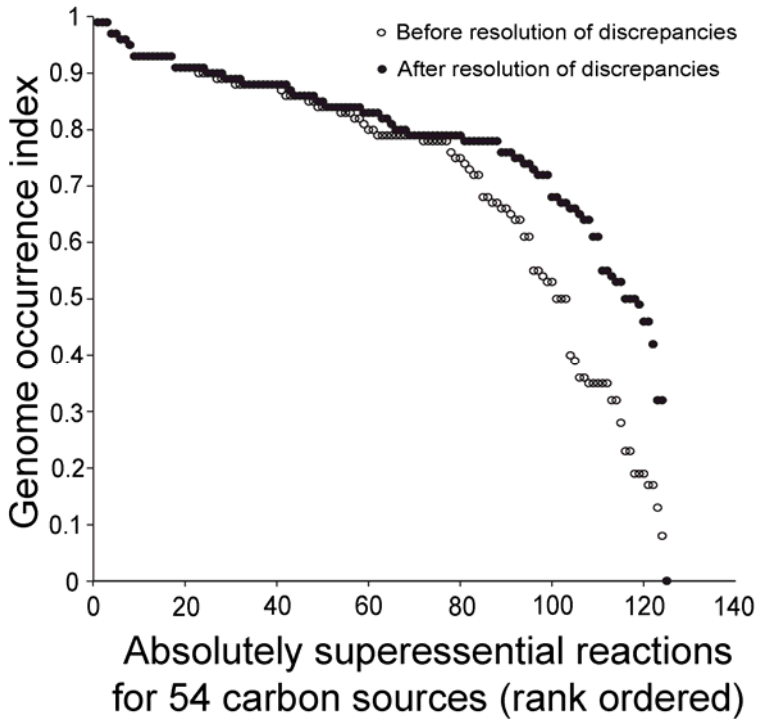


Figure 2.6: The plot shows the change in the genome occurrence index of 125 environment-general and absolutely superessential reactions when we manually improve the functional annotation of those 20 reactions that occur in fewer than 50 percent of genomes. Before the improvement, 105 reactions of the 125 reactions have a genome occurrence of more than 50 percent (○), while after improvement this number increases to 118 (94.4 percent) (●). See main text and supplementary material S4 in section 2.6 for details.

In sum, missing information about relevant enzymes and genes can lead to low apparent genome counts even for reactions with high superessentiality. We identified the reasons for low genome occurrence for those 20 of the 125 environment-general reactions in the absolutely superessential core that occur in fewer than 50 percent of genomes. Figure 2.6 shows a comparison of the genome occurrences before and after a correction for such discrepancies based on (limited) independent information. After correction, 94.4 percent (118 of 125) of absolutely superessential reactions occur in more than 50 percent genomes. The reasons for these discrepancies are similar to those listed in the above examples. They include misleading assignments of orthology, yet undiscovered enzymes, and non-orthologous gene displacement [42–46] (supplementary results in section 2.6 and supplementary material S4 in [29]).

Most absolutely superessential environment-general reactions remain superessential in complex *in vivo* and rich environments

So far, we have used 54 minimal environments distinguished by their sole carbon source to characterize reaction superessentiality. While the use of minimal environments makes our analysis simpler, it raises the question to what extent reaction superessentiality would be similar in the complex chemical environments that many pathogens need to survive. To answer this question, we used the universal network approach to identify absolutely superessential reactions in the complex environments known to sustain *in vivo* growth of *Salmonella typhimirium LT2* [47], *Mycobacterium tuberculosis H37Rv* [48], *Pseudomonas aeruginosa PAO1* [49], *Mycoplasma pneumoniae* [50], as well as a synthetic complete medium [51] (see methods, section 2.5). These environments consist of various nutrients such as amino acids, cofactors, fatty acids and nucleotides. In addition, we supplemented each environment with all 54 carbon sources we studied here, to render our identification of absolutely superessential reactions conservative, because additional nutrients will lead to a reduction, never to an increase of reaction superessentiality. Nonetheless, we found that in each of the five supplemented environments, a majority (at least 77.6 percent) of the 125 absolutely superessential reactions that we had previously identified (supplementary material S3 in [29]) are still superessential (supplementary material S5 in [29]). Among the absolutely superessential reactions that this approach identified (101 reactions on average for the five environments every single one is contained in our previously identified set of 125 absolutely superessential environment-general reactions. Furthermore, about 83 percent of these 101 reactions are represented in more than 50 percent of prokaryotic genomes (p-value < 10^{-5}, n= 10^5 for each of the five environments) (supplementary results and supplementary table S4 in section 2.6). The latter observation indicates not only the importance of these reactions. It also indicates that this importance is not restricted to organisms with a particular biomass composition such as that of *E. coli*, because the organisms in which these reactions occur are taxonomically diverse, and may vary widely in their biomass composition. To validate this assertion, that is, that the superessentiality of reactions is not highly sensitive to biomass composition, we also studied reaction superessentiality for different biomass compositions (supplementary results, section 2.6). We found that the relative magnitude of superessentiality indices is highly correlated between

different biomass compositions (Spearman's r = 0.8, *p*-value < 10^{-13}, *n* = 1561) (supplementary results, section 2.6).

2.4 Discussion

The metabolism of an organism can evolve through elimination of reactions due to loss-of-function mutations, and through addition of new reactions by horizontal gene transfer. Several studies, experimental and computational alike, have identified essential reactions [3–6,11–14,52] in different organisms or in genome-scale metabolic networks. Most such studies are organism-specific, and they do not address the question how readily a reaction could be by-passed through alternate reactions or pathways, based on our current knowledge of metabolism. This question is important not only to understand the evolutionary plasticity of metabolic networks, but also to identify those enzymatic targets for antimetabolic drugs where the risk of evolving resistance is smallest. Our approach of universal network analysis allows us to identify absolutely superessential reactions that are essential in any metabolic network. Random viable network sampling, in addition, allows us to quantify superessentiality for many reactions, and to study its causes in individual networks. Both approaches are complementary and can be used cross-validate each other.

One might argue that an analysis like ours should ideally use many reconstructed metabolic networks from diverse prokaryotes [53] instead of random viable network samples. However, the number of high-quality reconstructed networks suitable for FBA is currently still too low. In addition, all sets of such networks would be related through a common evolutionary history, which creates a bias in data that random viable network samples can avoid. Finally and relatedly, random viable network samples can be directly used for statistical hypothesis testing [22,23,54]. For our analysis, random viable metabolic networks are thus currently an indispensable tool.

We here focused on metabolic networks whose size is identical to that of *E.coli*, and whose viability is defined as the ability to synthesize all *E.coli* biomass precursors.

We did so, because the *E. coli* biomass composition is well-studied, and many of its components -- amino acids, nucleotides, etc. -- are representative of biomass precursors in most other free-living organisms. Moreover, *E. coli* is an environmental generalist, and thus able to survive in multiple different environments. This feature allowed us to compare reaction superessentiality for networks viable in one and in multiple environments. In this regard, we focused on minimal environments that vary in their sole carbon sources, because carbon is life's most central chemical element. Specifically, we compare networks viable in a minimal environment with glucose as its sole carbon source to networks that are viable on 54 different sole carbon sources. We refer to these two types of networks as networks with multiple and single carbon source phenotypes.

We began by identifying a core of absolutely superessential reactions (supplementary material S1 and S3 in [29]). These are reactions that occur and are essential in all networks we study. This core comprises 133 reactions for the single carbon source phenotype and 148 reactions in the multiple carbon source phenotype. The vast majority of reactions in this core are not specific to a given carbon source, but they are required for viability on all carbon sources. They also form a statistically highly significant connected component in a graph-based representation of metabolism. The properties of this core show that the reactions that are most difficult to by-pass do not allow the organism to survive in specific environments, but they are essential to life in multiple environments. Computational and experimental studies have tried to identify common essential reactions across a small number of organisms towards developing antibiotics [38,55]. Our identification of an absolutely superessential core of reactions goes beyond these analyses. It carries important implications for drug target identification. It predicts that antimetabolic drugs targeted towards enzymes that are most difficult to by-pass will not come from pathways that mediate adaptation to specific environments, but rather from anabolic pathways responsible for the synthesis of cell wall components and amino acids.

The next step of our analysis focused on the superessentiality of reactions for single and multiple carbon source phenotypes. We showed that reaction superessentiality in

single minimal and multiple minimal environments is highly correlated. This demonstrated again that a reaction's superessentiality derives mostly from reactions not specific to an environment. Second, we analyzed the relationship between a reaction's superessentiality and how frequently genes encoding known enzymes for this reaction occur in 1093 prokaryotic genomes. 94.4 percent (118 of 125) reactions in the environment-general superessential core occur in more than 50 percent of prokaryotic genomes, a number much greater than expected by chance alone. In addition, the statistical association between the superessentiality of a reaction and the number of genomes encoding it is much higher than expected by chance alone. Not unexpectedly, this association explains only a modest fraction of the variance in genome occurrence, which reflects our incomplete knowledge about metabolism and cell biology. For example, a highly superessential reaction catalyzed by several non-orthologous enzymes may show a low genome count, if genes encoding some of these enzymes have not yet been identified [42–44]. Other reasons for high superessentiality and low genome count involve promiscuous enzymes that catalyze more than one reaction, but are not known to do so [45,46], or unknown biochemical pathways that can by-pass a reaction [46]. Conversely, an enzyme with high genome count and low superessentiality may have important non-metabolic functions. Examples we discuss in the supplementary results (section 2.6) include the glycolytic enzyme enolase for its role in the RNA degradasome, or thioredoxin reductase for its indirect yet essential role in reducing important cytoplasmic enzymes and regulatory proteins.

We also studied how rich environments and environments known to support the life of pathogens inside a host affect the complement of absolutely superessential reactions. We found that a majority of reactions from a set of 125 absolutely superessential reactions needed for life on 54 carbon sources remained absolutely superessential in these environments. This observation is consistent with our earlier observations that most superessential reactions are environment-general. Moreover, enzyme-coding genes of 114 of the 125 absolutely superessential reactions have experimentally been confirmed as essential in at least one of three well-studied organisms, namely *S. enterica* serovars [38,56], *M. tuberculosis H37Rv* [57,58] and *P. aeruginosa* [59,60] (supplementary material S6 in [29]). Furthermore, a significantly larger number of absolutely superessential reactions than expected by chance alone (mean = 83

percent) are encoded by a majority of sequenced prokaryotic genomes. This indicates that these reactions are not just highly superessential because they support synthesis of biomass molecules highly specific to one organism, such as *E.coli*. All of the above means that the superessential core is not highly sensitive to the chemical composition of an environment and of biomass.

As knowledge about metabolism accumulates, our estimates of reaction superessentiality will become ever more accurate. Already at the present time, the estimates we obtained can help guide the selection of drugs targeting metabolism. That is, although our observations do not answer the question *how to inhibit* a specific enzyme, they can help answer *which essential enzymes* are the best candidates for inhibition, based on how difficult it is to bypass the reactions these enzymes catalyze. The main idea is that enzymes whose reactions have high superessentiality (large I_{SE}) are good drug target candidates, because cells cannot easily evolve resistance by re-routing metabolism around these enzymes [61]. This holds even more so for reactions with high superessentiality and high genome-count. One example is methionine adenosyltransferase (*metK*, Blattner number b2942), an absolutely superessential enzyme present in 97 percent of prokaryotic genomes. Other examples include shikimate kinase ($I_{SE} = 1$, $I_{GO} = 0.86$), and chorismate synthase ($I_{SE} = 1$, $I_{GO} = 0.89$), enzymes that are already being explored as possible targets [62,63]. As a further example, figure 2.7 shows a more detailed analysis involving the diaminopimelate (DAP) pathway. The DAP pathway is responsible for meso-2,6-Diaminopimelate (*m*-DAP) synthesis. *m*-DAP is an important precursor to L-lysine and peptidoglycan synthesis in prokaryotes, both of which are not synthesized in humans. (This is important as the ideal drug targets the pathogen but not the host.) DAP epimerase (E.C 5.1.1.7), the final enzyme involved in the production of *m*-DAP is actively being explored as a drug target [64–67]. Although DAP epimerase is essential for growth on all 54 carbon sources in *E.coli*, it is essential only in 54.2 percent of random viable networks, because many microbes can synthesize *m*-DAP through diaminopimelate

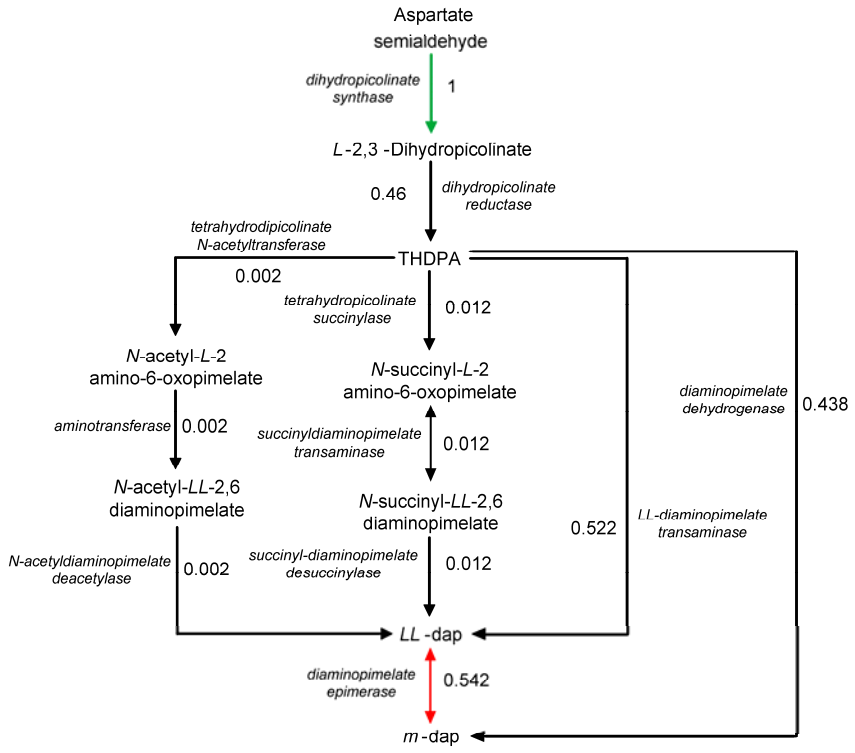


Figure 2.7: Diaminopimelate epimerase (red), a drug target that is being actively pursued, has an intermediate superessentiality index of because of an alternate pathway catalyzed by diaminopimelate dehydrogenase that also synthesizes *m*-DAP. In contrast, dihydropicolinate synthase (green) is absolutely superessential, making it a very attractive drug target. Note that the majority of the reactions shown possess alternate metabolic routes signified by their low superessentiality. THDPA: *L*-2,3,4,5-tetradihydropicolinate; *LL*-dap: *LL*-2,6-diaminopimelate; *m*-DAP: *meso*-2,6-diaminopimelate.

dehydrogenase (E.C 1.4.1.16) (reviewed in [68]). In other words, inhibition of this enzyme could be ineffective in the long run, because there is a known route of resistance. A better target in the same pathway would be dihydrodipicolinate synthase (E.C 4.2.1.52), which is superessential, environment-general, and present in 92 percent of prokaryotic genomes. In addition, its product, *L*-2,3-dihydropicolinate is not essential in humans, making dihydrodipicolinate synthase an ideal drug target. In sum, the analysis of superessential reactions may be one of several worthy starting points towards development of drug targets that block or slow down the evolution of antidrug-resistance [69–71].

We next discuss potential caveats to our study. First, as noted above, estimation of reaction superessentiality depends on our knowledge about the "universe" of all feasible enzyme-catalyzed metabolic reactions. If future work adds reactions and pathways to the known universe, then the superessentiality of individual reactions may decline. However, we expect that the ranking of superessentiality of many reactions will remain similar over time. If so, a reaction whose superessentiality is much higher than that of another reaction would still be a better candidate drug target. Second, our comparison of essential reactions to enzyme occurrence in organisms with sequenced genomes depends on the quality of available metabolic genome annotation [28,37]. This annotation currently has numerous limitations, as we discussed. Third and relatedly, completely unknown spontaneous (non-enzyme-catalyzed) reactions could lower the superessentiality of enzyme-catalyzed reactions. They could thus contribute to some of the yet unresolved low genome occurrences of absolutely superessential reactions. However, based on our knowledge of known spontaneous reactions, which constitute a very small fraction (1.3 percent, supplementary results, section 2.6) of all known reactions, this effect may be minor. A fourth caveat regards uncertainties in biomass composition. We carried out our analysis with the biomass composition of a free-living organism in mind, and the majority of biomass molecules we consider would be found in typical free-living organisms. However, some minor biomass components may be restricted to some organisms and may not be required in others. A potential example regards siroheme, a molecule that is part of *E.coli* biomass but may not be needed in other organisms [47,53,72–74]. In an organism that does not need this molecule 1.6 percent (two reactions) of our absolutely superessential reactions lose this status. However, even when taking some variation in biomass composition into account, the relative order of superessentiality among reactions would be largely preserved. To show this, we used the biomass composition of SEED models, metabolic models of diverse organisms that are created with a semi-automatic procedure from genomic and other information [53]. Specifically, we re-calculated the superessentiality index (I_{SE}) of reactions for a biomass composition that contains molecules found in a majority of these SEED models. The rank correlation coefficient between superessentiality indices for the *E.coli* biomass and for this biomass composition exceeds 0.8. This means that the superessentiality index is not highly sensitive to biomass composition. Lastly, the generation of random viable networks,

especially for multiple carbon source phenotypes is computationally expensive [22,75]. Only limited sample sizes are currently feasible. However, we note that even these limited sample sizes yield results that are in agreement with complementary approaches, for example the identification of absolutely superessential reactions from the "universal" network that we analyzed.

In sum, our analysis sheds light on the evolution and genotypic plasticity of metabolic networks. It shows that metabolic networks contain a core of absolutely superessential reactions regardless of their metabolic genotype. The composition of this core is not highly sensitivity to the environment in which viability is required. More generally, reactions vary broadly in their superessentiality, and thus in how readily they can be by-passed by alternative pathways. A comprehensive evolutionary approach like ours may help identify putative drug targets and to develop effective antibiotic therapies. We hope that our data on reaction superessentiality (supplementary materials S1, S3, S4 and S5) will become a broadly useful resource to other researchers.

2.5 Methods

Flux Balance analysis

Flux balance analysis (FBA) is a constraint-based modeling method which uses information about enzymatic reactions and the stoichiometry of each reaction to predict steady state fluxes for all reactions in a metabolic network [75,76]. In addition, it also predicts the maximum biomass yield that a metabolic network can achieve [75,76]. Specifically, FBA uses the stoichiometric information of all reactions in a network, as given by the stoichiometric matrix **S** [76]**.** This matrix has dimensions *m x n*, where *m* denotes the number of metabolites and *n* denotes the number of reactions in a network. FBA assumes steady state conditions, where one can impose the constraint of mass conservation on the metabolites in the network. This constraint can be written as

$$\mathbf{Sv} = 0$$

wherein **v** denotes the vector of metabolic fluxes through each network reaction. For genome-scale metabolic models, the above equation leads to an under-determined system with a large solution space of allowable fluxes. Further constraints based on regulatory and thermodynamic information can be used to reduce the number and limit the magnitude of fluxes within the network. Linear programming can then be used to identify a set of allowable flux values that maximize a biologically relevant objective function ***Z*** [76]. The linear programming formulation of an FBA problem can be written as:

$$\max \mathbf{Z} = \max \{\mathbf{c}^T\mathbf{v} \mid \mathbf{S}\mathbf{v} = 0, \mathbf{a} \leq \mathbf{v} \leq \mathbf{b}\}$$

wherein the vector **c** contains a set of scalar coefficients that encapsulate the maximization criterion, and vectors **a** and **b** respectively contain lower and upper bounds for each reaction in **v**.

We are here interested in predicting whether a particular network can sustain life in a given environment, which means that the network can synthesize all biochemical precursors necessary to sustain growth and energy production. We use the *E.coli* biomass composition from the *E. coli* metabolic model *i*AF1260 [12] to define the vector **c**. We used the packages CPLEX (11.0, ILOG; http://www.ilog.com/) and CLP (1.4, Coin-OR; https://projects/coin-or.org/Clp) to solve the linear programming problems mentioned above.

Growth environments

In addition to the stoichiometric matrix and the biomass objective function, one needs to define a chemical environment that contains the different nutrients needed to synthesize biomass precursors. We here consider only minimal growth environments composed of one carbon source, along with oxygen, ammonium, inorganic phosphate, sulfate, sodium, potassium, cobalt, iron (Fe^{2+} and Fe^{3+}), protons, water, molybdate, copper, calcium, chloride, magnesium, mangnese and zinc [12]. When studying viability of a metabolic network on different minimal environments, we vary different carbon sources while keeping all other nutrients constant. For instance, if a particular

network is viable in 54 carbon sources, we mean that the network can synthesize all essential biochemical precursors when each of these carbon sources is provided as the sole carbon source in a minimal medium. We restrict ourselves to the 54 carbon sources that *E. coli* is known from experiment to be viable in [12], listed in supplementary table 1 in section 2.6.

We have also used complex and *in vivo* environments to identify essential reactions from the universal network (see below for *essential reactions* and *universal network*). Specifically, we used environments characterized for in-vivo growth of *Salmonella typhimirium LT2* [47], *Mycobacterium tuberculosis H37Rv* [48], *Pseudomonas aeruginosa PAO1* [49], *Mycoplasma pneumoniae* [50] and a laboratory growth medium called the synthetic complete medium [51]. We supplemented all five environments mentioned above with all 54 carbon sources (supplementary table 1, section 2.6) to render our identification of absolutely superessential and environment-general reactions conservative, because addition of nutrients can only lead to a reduction of a reaction's superessentiality.

Genotypes, phenotypes and viability

A metabolic network is comprised of reactions that allow it to synthesize all biomass precursors from a given set of nutrients. We define the complement of reactions that take place in a network as the *metabolic genotype* of that network. Any one organism's genotype exists in a much larger space of metabolic genotypes. This space is defined by the biochemical reactions known to be realized in living cells. Any one organism's genotype can be thought of as a point in this space, where some biochemical reactions occur and some do not. Two genotypes differing from each other by one reaction are thus neighbors in the genotype space.

We define the *phenotype* of such a genotype as its ability to sustain life in a given minimal environment or a set of environments. Specifically, we consider a genotype to be *viable* in a given environment only if its maximum biomass flux as predicted by FBA is nonzero. We thus represent a metabolic phenotype as a binary string, whose *i*-th entry is one if a network is able to sustain life when C_i is the sole carbon source.

universe of reactions is a list of metabolic reactions known to occur in some organism. For the construction of this universe, we used data from the LIGAND database [27,28] of the Kyoto Encyclopedia of Genes and Genomes (37, 73). Briefly, the LIGAND database is divided into two subsets – the REACTION and the COMPOUND database. Together, they incorporate data on reactions, associated stoichiometric information and the chemical compounds involved therein in an interlinked fashion. Also included in the REACTION database is the Enzyme Classification (E.C.) identifier of each reaction, as provided by the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB) (http://www.chem.qmul.ac.uk/iubmb/enzyme/).

We specifically used the REACTION and the COMPOUND databases to construct our universe of reactions while excluding - (i) all reactions involving polymer metabolites of unspecified numbers of monomers, or general polymerization reactions with uncertain stoichiometry, (ii) reactions involving glycans, due to their complex structure, (iii). reactions with unbalanced stoichiometry, and, (iv) reactions involving complex metabolites without chemical information about their structure [22].

The published *E. coli* metabolic model (*i*AF1260) consists of 1397 non-transport reactions [12]. We merged all reactions in the *E. coli* model with the reactions in the KEGG dataset, and retained only the unique (non-duplicate) reactions. After these procedures of pruning and merging, our universe of reactions consisted of 5906 non-transport reactions and 5030 metabolites.

Identification of blocked reactions

Well curated genome-scale metabolic models can contain reactions that have a zero flux in a given growth environment under any steady-state condition with non-zero biomass growth flux [12]. Such reactions have been considered to be "blocked" [33]. We identified blocked reactions by maximizing the flux through each reaction in the

network and considered a reaction blocked in a given environment if no positive flux ($< 10^{-8}$) was achievable through it.

Generation of random viable metabolic networks

We here consider evolutionary change in metabolic genotypes through the addition of enzyme-coding genes or through loss of such genes. Over large evolutionary time scales, the reactions in a network may change dramatically, while natural selection may play an important role in preserving the viability of the organism in its original environment. We here employ a previously described *in silico* evolutionary process referred to as a Markov Chain Monte Carlo (MCMC) random walk to generate networks comprising random metabolic reactions that are viable in the specified carbon sources [22,23].

Each step in the random walks we use can be either the addition or deletion of a reaction. When adding a reaction, we add a randomly chosen reaction from the universe of reactions. Because our approach requires unbiased random walks, we need to consider all reactions in the universe of reaction for addition, which can result in the addition of blocked reactions in the mutating network. Addition of any reaction does not eliminate the viability of a network in any one environment. We always follow a reaction addition by a deletion of a randomly chosen reaction. After a reaction deletion, we use FBA to predict the viability of a network within one or more chemical environments. We accept the reaction deletion if the network remained viable after deletion; otherwise we reject the deletion and another reaction is randomly chosen for deletion. We call the process of sequential addition and deletion of reactions a reaction swap. We use such reactions swaps, because they enable us to keep the number of reactions in a network constant throughout each random walk. Transport reactions are not subject to addition or deletion, and thus are always present in all networks generated by our random walks.

For the MCMC method to produce random samples of networks, it is important to carry out as many reaction swaps as necessary to “erase” the similarity to the initial network. Successive genotypes in the Markov chains are strongly correlated, since

they differ by one reaction pair, though this "memory" fades with the number of steps (reaction swaps) performed. These correlations, which can be described through an autocorrelation function that can be estimated empirically, decrease exponentially with the number of swaps. For network sizes of 1397 reactions, we found that 3 x 10^3 reaction swaps suffice to erase the similarity between the initial and the final network [23,78]. Thus, a network at the end of 3x10^3 swaps is essentially randomized with respect to the starting network. We refer to it as a *random viable metabolic network*. To err on the side of caution, we allowed our Markov chains to run for even more (5 x 10^3) swaps before sampling a network. These observations motivate the following procedure to generate multiple random viable networks. We start a random walk from the *E. coli* metabolic network, continue this walk for a total of 2.5 x 10^6 reaction swaps, and sample a viable metabolic network every 5 x 10^3 swaps. In this fashion, we generate 500 random viable metabolic networks with the same size as *E. coli* network.

Successive genotypes in an MCMC random walk are connected in the genotype space through one reaction swap. This also means that two genotypes at extreme ends of the Markov Chain are connected through a series of such swaps. One can thus effectively travel from one genotype to a farther genotype with *a priori* information of the mutations involved therein. Thus, all networks generated by our method are "connected" to each other and lie on the same genotype network. As we start our random walk with the *E. coli* metabolic network [12], it is important to note that our method of sampling applies only to the "connected component" of the genotype network, of which the *E. coli* metabolic network is one of the nodes.

Identification of essential reactions

We define a reaction as essential for a given genotype if its elimination abolishes the network's ability to synthesize all biomass precursors in a specific chemical environment. To identify such essential reactions in a given network, we eliminated each reaction and used FBA to assess whether non-zero biomass growth flux was still achievable. If not, we referred to a reaction as *essential* for this genotype and growth environment. We called a reaction *absolutely superessential* if it was essential in all random viable metabolic networks in a given environment.

Universal network

We converted the known universe of reactions into a universal metabolic network by including the *E. coli* transport reactions in it [12]. We note that the universal network may not be biologically realizable. For example, it may contain reactions or pathways that are thermodynamically infeasible. However, it serves as a useful reference point in our work, where it can be used to assess whether the limited sample sizes of random viable networks that we can generate affect the validity of our results. They do not, as we discuss in the main text.

Estimation of connected components in reaction subgraphs

To test the null hypothesis that the set of 125 environment-general, absolutely superessential reactions (superessential core) would be expected to form a connected network, we first represented the universe of reactions as a reaction graph. In such a graph, reactions are represented as vertices that are linked by edges if a pair of reactions shares a common metabolite. We then extracted the reaction subgraph which corresponded to the superessential core and determined the number of strongly connected components and the size of the largest component using the Matlab boost-graph library by David Gleich (http://www.stanford.edu/~dgleich/programs/matlab_bgl/ and http://www.mathworks.com/matlabcentral/fileexchange/10922). We then randomly selected 125 reactions (equal to number of reactions in the superessential core) from the complete reaction graph, and determined the number of components and the size of the largest component of the subgraph comprising these reactions. We repeated this randomization procedure 10^5 times. The probability (p-value) to find the observed number of components in the superessential core by chance alone is given by the fraction of these 10^5 trials in which that yielded more connected components than the superessential core. The p-value for the size of the largest component calculates analogously.

We repeated the above procedure under the exclusion of currency metabolites. Currency metabolites transfer small chemical groups and are involved in a large

number of reactions [79]. The currency metabolites we eliminated are ATP (adenosine triphosphate), ADP (adenosine diphosphate), AMP (adenosine monophosphate), water (H_2O) and protons (H^+). To eliminate them, we neglected all edges emanating from these metabolites when constructing the reaction graph.

Pathway enrichment

As discussed in the main text, we found 133 absolutely superessential reactions for growth on glucose and 125 reactions absolutely superessential and environment-general for growth on 54 alternative carbon sources. To ask whether reactions in this core preferentially derive from specific pathways, we performed the following analysis for each metabolic pathway *P*. We used the existing classification of all 1397 *E. coli* reactions into different pathways [12] to calculate the fraction of these 1397 reactions that are associated with *P*. This fraction serves as a null-hypothesis about the expected fraction of reactions in *P*. Subsequently, we determined the number of reactions in our superessential core (133 and 125 reactions for growth on glucose and on 54 alternative carbon sources respectively) that belong in *P*. We then used an exact binomial test to ask whether the fraction of reactions from the superessential core that belong in *P* is significantly different from the expected fraction of reactions in *P*. We corrected the resulting *p*-values from the binomial test for multiple hypotheses testing using the Benjamini-Hochberg false discovery rate [80].

Genome occurrence index

To test the hypothesis that reaction superessentiality is correlated with the number of prokaryotic genomes that contain the corresponding enzyme-coding gene we used data from the KEGG GENES database. The GENES database holds information on annotated genes from a total of 1093 prokaryotic organisms. Each gene from an organism is assigned to an orthologous group which can be classified based on the KEGG Orthology (KO) number (http://www.genome.jp/kegg/ko.html), and which can help estimate the number of genomes that contain a gene orthologous to any gene of interest. For *E. coli*, this list contains KO assignments for 3004 genes (ftp://ftp.genome.jp/pub/kegg/genes/organisms/eco/eco_ko.list) as of December 2010. The published *E. coli i*AF1260 model lists gene annotations for 1295 out of the 1397 non-transport reactions, enabling us to assign KO numbers to about 93 percent of *E.*

coli reactions in the known reaction universe. For all other reactions in the known reaction universe, we directly assigned a KO number.

Assessing the significance of reactions possessing absolute superessentiality (I_{SE} = 1) and reactions possessing high genome-occurrence index ($I_{GO} \geq 0.5$)

To test the null hypothesis that absolutely superessential reactions occur in a majority of genomes just by chance, we needed to determine the probability that a randomly chosen reaction from the list of essential reactions has a high genome occurrence index ($I_{GO} \geq 0.5$). Of the 1569 reactions that are essential for growth on 54 carbon sources in at least one network (figure 2.6), 155 reactions are absolutely superessential. From the list of absolutely superessential reactions, we calculated the number of reactions which occur in 50 percent or more genomes ($I_{GO} \geq 0.5$). We found 114 such reactions. We then chose 10^5 random subsets of 155 reactions from this list of 1569 essential reactions. We calculated the *p*-value via the fraction of samples in which the number of reactions with high genome occurrence ($I_{GO} \geq 0.5$) was at least 114.

We calculated similar statistics for occurrence of absolutely superessential reactions in more than 50 percent prokaryotic genomes for growth in complex and *in-vivo* environments as explained above. Instead of using 114 reactions from a core of 155 reactions to test for the null hypothesis, we used the exact number of reactions found to occur in more than 50 percent prokaryotic genomes from a core of 125 reactions. This exact number is given in supplementary results (section 2.6) for each complex medium.

Assessing the significance of correlation between superessentiality index (I_{SE}) and genome-occurrence index (I_{GO}) for all reactions

To determine whether the Spearman's rank correlation coefficient of $\rho = 0.4$ between the superessentiality index and the genome occurrence index could have arisen chance alone, we proceeded as follows. We viewed the data for these indices as two vectors, randomly permuted the entries of the vector of superessentiality indices, and calculated ρ again from the permuted data. We repeated this randomization 10^5 times,

and determined our probability of interest (*p*-value) as the fraction of these randomizations where ρ was no smaller than in the actual data.

All numerical analyses described here were carried out using Matlab (Mathworks Inc.).

2.6 Supplementary results

The environment-general superessential core forms a compact, highly connected component

As discussed in the main text, we found 125 reactions which were absolutely superessential and environment-general. We referred to this set of reactions as the superessential core. We asked whether the reactions in this superessential core come from different, far-flung regions of the universal metabolic network, or whether they occur in a highly localized part of this network. We addressed this question by representing the universal reaction network as a graph (see methods, section 2.5) [81], and examined the connectedness of its subgraphs. We found that the superessential core of 125 environment-general reactions forms two connected subgraphs, one comprising 124 reactions, and a "trivial" subgraph consisting of a single reaction disconnected from this subgraph. We then sampled 10^5 sets of 125 reactions from the universal network, and studied the distribution of their number of connected subgraphs, as well as the size of their largest component (supplementary figures S2A and S2B). This analysis showed that the existence of only two components is highly unlikely to occur by chance alone (p-value < 10^{-5}, $n = 10^5$, randomization test), and that a largest connected component of size 124 reactions is also highly unlikely to occur by chance alone (p-value < 10^{-5}, $n = 10^5$). Thus, the superessential core consists of a compact and localized subset of metabolism. We note that the isolated reaction in the superessential core is periplasmic in nature and thus disconnected from the rest of metabolism by its very nature. This reaction is catalyzed by a murein crosslinking transpeptidase and generates a peptidoglycan-related biomass component. We repeated this analysis while excluding highly connected currency metabolites (see methods, section 2.5), and found that the number of components in the superessential core increased from two to six. Similarly, the size of the largest component decreased

from 124 to 119. However, the number of connected subgraphs was still significantly smaller than expected by chance (p-value < 10^{-5}, $n = 10^5$), and their largest component is still significantly larger (p-value < 10^{-5}, $n = 10^5$).

Some reactions have high superessentiality in multiple carbon source environments and low superessentiality in single carbon source environments

In this section, we investigate the small number of reactions whose superessentiality index is very different in single carbon source and multiple carbon source environments (figure 2.4). Specifically, we want to answer how some reactions can be essential in many networks (high I_{SE}) for multiple carbon source phenotypes, but non-essential in many networks (low I_{SE}) in the glucose phenotypes. The answer can be understood from an extreme example, the reaction catalyzed by thymidine phosphorylase (TMDPP), which is a part of pyrimidine metabolism. This reaction is absolutely superessential ($I_{SE} = 1$) for viability on 54 carbon sources, but it has an I_{SE} of only 0.004 in networks viable on glucose (figure 2.4). The enzyme catalyzing this reaction is non-essential for growth on glucose and glycerol in *E.coli* [11,15]. The substantial difference in superessentiality we observe for this reaction arises because the reaction is absolutely superessential for growth in thymidine, one of our 54 carbon sources. It is thus also absolutely essential for the multiple carbon source phenotypes we consider. (Not surprisingly, the reaction is essential for growth on thymidine in *E. coli*). This explanation highlights a more general theme: Some reactions that are not very important for growth on glucose can have high superessentiality in a multiple carbon source phenotype, namely by being hard or impossible to by-pass when growth in a specific environment is required. Other examples of such reactions include deoxyadenosine utilizing purine-nucleoside phosphorylase, which is often required for growth on deoxyadenosine ($I_{SE} = 0.768$ on 54 carbon sources), and (deoxyribose) phosphopetomutase 2, which is often required for growth on deoxyadenosine and thymidine ($I_{SE} = 0.862$ on 54 carbon sources).

Absolutely superessential reactions occur in a large number of prokaryotic genomes

If a reaction is frequently essential for preserving a phenotype in random viable

metabolic networks, then the corresponding enzyme-coding gene should also occur frequently in many prokaryotic genomes. This line of reasoning will fail if either our knowledge of the universe of metabolic reactions is incomplete, or if our understanding of an organism's enzyme complement is partial. The extent to which it fails can shed light on how imperfect our current metabolic knowledge is. We therefore turn to some examples which illustrate the sources of this incompleteness, which reside in metabolic gene annotation [82]. First, discrepancies may arise due to misleading assignment of orthologous groups. For instance, anthranilate phosphoribosyltransferase (ANPRT) is an absolutely superessential enzyme necessary for tryptophan biosynthesis. ANPRT is coded by the gene *trpD* (Blattner number b1263) [83] and occurs in 84 prokaryotic genomes ($I_{GO} = 0.077$) according to current orthology assignments. This discrepancy between the absolute superessentiality and low genome occurrence occurs because the orthology of *trpD* is based on a different enzyme called anthranilate synthase, a heterologous enzyme encoded by products of *trpD* and another gene *trpE* [83]. Instead, orthology of *trpD* when based on anthranilate phosphoribosyltransferase activity alone revealed that this reaction occurs in 853 genomes ($I_{GO} = 0.78$). Similar examples exist where identical reaction-catalyzing enzymes are encoded by non-orthologous genes [42–44], thus reinforcing the essentiality and occurrence of a reaction across many organisms.

Many further reactions show high superessentiality but apparently low genome occurrence because they are needed for growth in a very limited number of carbon sources. For instance, the enzyme allulose-6-phosphate epimerase (ALLULPE) carries out the final step in the degradation of D-allose [84], one of the 54 carbon sources analyzed here. Elimination of ALLULPE abolishes growth only on D-allose in all random metabolic networks and in *E. coli*. Thus, organisms that do not use D-allose as a carbon source do not need the gene for encoding this enzyme.

The opposite kind of discrepancy regards reactions with a low superessentiality index (I_{SE}) yet a high genome occurrence index (I_{GO}). We found that this kind of discrepancy may arise when a protein acts not only as an enzyme, but possesses other non-enzymatic functions that may be important, but that are not captured by a

metabolic analysis. For example, the enzyme enolase (*eno*, Blattner number b2779), which carries out a reversible conversion between 2-phosphoglycerate and phosphoenolpyruvate has a low superessentiality of $I_{SE} = 0.134$, while being encoded by many genomes ($I_{GO} = 0.98$). The reason is that apart from its glycolytic function, enolase also plays an important role in the RNA degradasome [85,86], an essential function outside small molecule metabolism. Similarly, NADPH-dependent thioredoxin reductase *(trxB,* Blattner number b0888) (TRDR) reduces thioredoxin in *E.coli*. While TRDR itself does not take part in many metabolic processes, its substrate thioredoxin is essential in reducing important cytoplasmic enzymes and regulatory protein [87,88]. This role is independent of small molecule metabolism and may explain the encoding gene's occurrence in many genomes. Multiple other reactions with similar discrepancies between superessentiality and genome occurrence are discussed in the supplementary material S4 in [29].

Most absolutely superessential environment-general reactions remain superessential in complex *in-vivo* and rich environments

We know that growth environments with different nutrients may affect the essentiality of a reaction. Pathogens residing in the human body experience a relatively nutritious environment in comparison to single carbon source minimal environments. Such environments are characterized by the availability of different sugars, amino-acids, fatty acids and various cofactors [89]. Does the set of absolutely superessential environment-general reactions based on minimal environments change dramatically when one studies growth in *in vivo* complex environments?

To address this question, we first identified absolutely superessential reactions for a medium that contained all 54 carbon sources, but that was minimal for all other nutrients. We found that all such reactions identified through the universal network approach were identical to our superessential core consisting of 125 reactions. This is not surprising. Each reaction in the superessential core is environment-general, where the environments we studied differ from each other only in their carbon source. Combining all carbon sources into one medium is thus results in identification of the superessential core.

We next used the *in-vivo* environments characterized for the organisms *Salmonella typhimurium LT2* [47], *Mycobacterium tuberculosis H37Rv* [48], *Pseudomonas aeruginosa PAO1* [49], *Mycoplasma pneumoniae* [50]. In addition, we also analyzed the effect of synthetic complete medium [51], a generic well-defined growth medium, in an effort to analyze a complex medium for organisms with uncharacterized growth environments. We then identified absolutely superessential reactions via the universal network for each of the above environments. Because many of the nutrients in rich environments, such as amino-acids and cofactors are also biomass components, one may expect that the number of superessential reactions will decrease drastically. This is not the case. From our set of 125 reactions that form the superessential core, we found that between 97 reactions (77.6 percent) and 108 reactions (86.4 percent) were essential for growth in one of the *in vivo* environments (supplementary material S5 in [29] and supplementary table S4).

Each of the reactions identified in this analysis belongs to our set of 125 absolutely superessential environment-general reactions. Thus, a majority of reactions in the superessential core are still absolutely superessential in complex environments. Furthermore, at least 80 percent of enzyme-coding genes responsible for these reactions occur in more than 50 percent of prokaryotic genomes (p-value $< 10^{-5}$, $n= 10^5$) (supplementary table S4), underscoring the generality of our results for different biomass compositions that may be found in different organisms. Possible reasons for genes occurring in fewer than 50 percent genomes have been explained above in supplementary material S4 in [29].

Superessentiality of reactions is not highly sensitive to biomass composition

We carried out our analysis with the biomass composition of a free-living organism in mind. It is clear that synthesizing the biomass of an endosymbiont or endoparasite which typically receives multiple biomass molecules from its host would require fewer reactions. The majority of the biomass molecules we analyzed would be required in most free-living organism. They include amino acids, DNA and RNA nucleotides, cofactors such as pyridoxal–5-phosphate, flavin adenine mononucleotide,

riboflavin, coenzyme A, cell membrane and cell wall constituents such as undecaprenyl phosphate, phosphatidic acid-based molecules and lipopolysaccharides [12]. We hypothesized that even taking biomass variation into account, the relative order of the superessentiality index among reactions would not change much.

To assess how varying biomass composition affects the superessentiality index of reactions, we used the biomass compositions of 129 SEED models, which are genome-scale metabolic networks models created through a semi-automatic procedure from assembled genome sequences [53]. These set of models include the metabolisms of a diverse spectrum of organisms [53]. While some biomass molecules, such as amino acids, DNA, and RNA nucleotides are universally required in free-living organisms, others, such as some cofactors and lipids may vary. We varied the biomass to include all biomass molecules found in at least 90 (70 percent) of the 129 SEED models. The resulting biomass composition did not include any biomass precursors that did not also occur in *E.coli*, but only a subset of *E.coli* biomass molecules (Note that the requirement to synthesize fewer molecules can only decrease, not increase the superessentially index of a reaction). The reduced biomass composition consisted of 55 molecules as opposed to the 63 molecules of *E.coli*. Specifically, we removed heme, siroheme, protoheme, thiamin diphosphate, 2-octaprenyl-6-hydroxyphenol (quinone), 5-methyltetrahydrofolate, 10-formyltetrahydrofolate and phosphotidylethalomine. We then generated 500 random metabolic networks viable on glucose and recalculated the superessentiality index of all reactions essential for the new biomass composition. We found that of the 133 reactions which were absolutely superessential, 122 (92 percent) still remained absolutely superessential for growth on glucose. Furthermore, the rank correlation coefficient between the superessentiality indices of reactions for the original and the reduced biomass is high (Spearman's r = 0.8, p-value < 10^{-13}, n = 1561). This means that the relative order of superessentiality among reactions is largely preserved. Frequently essential reactions remain frequently essential even with varying biomass.

The effect of removal of spontaneous reactions on the absolute superessentiality of reactions

While most biochemical reactions are catalyzed by enzymes, some reactions proceed

spontaneously and do not require catalysis. Some of these spontaneous reactions may link substrates and products that can also be linked through one or more other enzyme-catalyzed reactions. Removal of such spontaneous reactions from the networks we study could convert non-essential reactions into absolutely superessential reactions. We wished to understand whether this hypothetical phenomenon would dramatically influence the number of superessential reactions we identify. To this end we identified all spontaneous reactions in the known reaction universe, which constitute only a tiny fraction of 1.3 percent (84 reactions) of all reactions in the universe. (This identification was based on reaction annotation in the *E. coli* metabolism [12] and in the KEGG reaction database [37,77]). We then removed all spontaneous reactions from our universal network, and determined whether doing so dramatically increased the total number of environment-general absolutely superessential reactions for growth on multiple carbon sources. The answer is no. Only a single candidate for an additional superessential reaction emerged, which is catalyzed by GTP cyclohydrolase (E.C. 3.5.4.16, gene name *folE*).

GTP cyclohydrolase catalyzes a kinetically complex first step in tetrahydrofolate synthesis. This reaction is represented by four distinct reactions in the KEGG reaction database, two of which are annotated as spontaneous therein. The same reaction is, however, represented as a one-step catalytic reaction in the *E. coli* metabolism [12]. We had removed reactions identical in the KEGG database and the *E. coli* model during the construction of the reaction universe (see methods, section 2.5), but because the KEGG reaction database includes a four step reaction instead of a single reaction, the reaction universe effectively contains two pathways representing the same overall reaction. Removal of the two spontaneous steps from the KEGG database resulted in the one-step reaction in *E. coli* to be absolutely superessential. However, closer inspection revealed even the one-step reaction to be ambiguous in its status. Specifically, it involves a series of four different chemical transformations. The last two of them may indeed be spontaneous [90], but the first two are enzyme catalyzed. Because of the problematic status of this reaction, we did not increase the number of reactions in our superessential core by the reaction in question, in order to err on the side of caution.

Supplementary table S1 – Names of the 54 carbon sources used in the study

5-Dehydro-D-gluconate	D-Glucarate
Acetate	D-Glucuronate
N-Acetyl-D-glucosamine	L-Glutamine
N-Acetyl-D-mannosamine	Glycine
N-Acetylneuraminate	Glycolate
Adenosine	Inosine
2-Oxoglutarate	L-Lactate
D-Alanine	Lactose
L-Alanine	L-Lyxose
D-Allose	D-Malate
L-Arabinose	L-Malate
L-Aspartate	Maltose
Deoxyadenosine	Maltotriose
D-Fructose 6-phosphate	D-Mannose
Formate	Melibiose
D-Fructose	D-Mannitol
L-Fucose	Pyruvate
Fumarate	D-Ribose
D-Glucose 1-phosphate	L-Rhamnose
D-Glucose 6-phosphate	D-Sorbitol
D-Galactose	D-Serine
D-Galactarate	L-Serine
D-Galactonate	Succinate
L-Galactonate	Thymidine
D-Galacturonate	Trehalose
D-Glucose	Uridine
D-Gluconate	D-Xylose

Supplementary table S2 – Pathway enrichment for 133 absolutely superessential reactions for growth on glucose

Pathway	Fraction of reactions in each pathway in the absolutely superessential core for glucose (133 reactions)	Fraction of reactions in each pathway for all *E. coli* reactions	Exact Binomial *p*-value	Benj. Hochberg corrected *p*-value	Number of abs. superessential reactions
Cell Envelope Biosynthesis	0.2707	0.0959	5.21E-009	5.21E-009	HTE
Purine and Pyrimidine Biosynthesis	0.0977	0.0172	5.13E-007	5.53E-007	HTE
Histidine Metabolism	0.0602	0.0072	5.73E-006	6.69E-006	HTE
Cofactor and Prosthetic Group Biosynthesis	0.2481	0.1160	1.09E-005	1.38E-005	HTE
Glycerophospholipid Metabolism	0.0752	0.1611	1.64E-003	2.30E-003	LTE
Valine, Leucine, and Isoleucine Metabolism	0.0451	0.0107	2.62E-003	4.08E-003	HTE
Tyrosine, Tryptophan, and Phenylalanine Metabolism	0.0526	0.0157	3.98E-003	6.97E-003	HTE

Nucleotide Salvage Pathway	0.0376	0.0938	7.81E-003	1.56E-002	LTE
Membrane Lipid Metabolism	0.0075	0.0301	7.09E-002	1.65E-001	-
Lipopolysaccharide Biosynthesis / Recycling	0.0677	0.0487	8.59E-002	2.40E-001	-
Threonine and Lysine Metabolism	0.0150	0.0129	2.67E-001	9.33E-001	-
Murein Biosynthesis	0.0075	0.0100	3.53E-001	3.53E-001	-
Glutamate metabolism	0.0075	0.0043	3.24E-001	2.27E+000	-
Methionine Metabolism	0.0075	0.0100	3.53E-001	4.94E+000	-

Significant *p*-values are highlighted in red (Benjamini-Hochberg test p-value < 0.05) [80]. Abbreviations: HTE – higher than expected; LTE – Lower than expected.

Supplementary table S3 – Pathway enrichment for 125 absolutely superessential environment-general reactions for growth on all 54 carbon sources.

Pathway	**Fraction of reactions in each pathway in the absolutely superessential core for glucose (125 reactions)**	**Fraction of reactions in each pathway for all *E. coli* reactions**	**Exact Binomial *p*-value**	**Benj. Hochberg corrected *p*-value**	**Number of abs. superessen-tial reactions**
Cell Envelope Biosynthesis	0.2800	0.0959	3.25E-009	3.25E-009	THE
Histidine Metabolism	0.0640	0.0072	3.65E-006	3.95E-006	THE
Cofactor and Prosthetic Group Biosynthesis	0.2640	0.1160	2.77E-006	3.27E-006	THE
Valine, Leucine, and Isoleucine Metabolism	0.0480	0.0107	1.96E-003	2.54E-003	THE
Tyrosine, Tryptophan, and Phenylalanine Metabolism	0.0560	0.0157	2.90E-003	4.19E-003	THE
Glycerophospholipid Metabolism	0.0800	0.1611	3.52E-003	5.72E-003	LTE

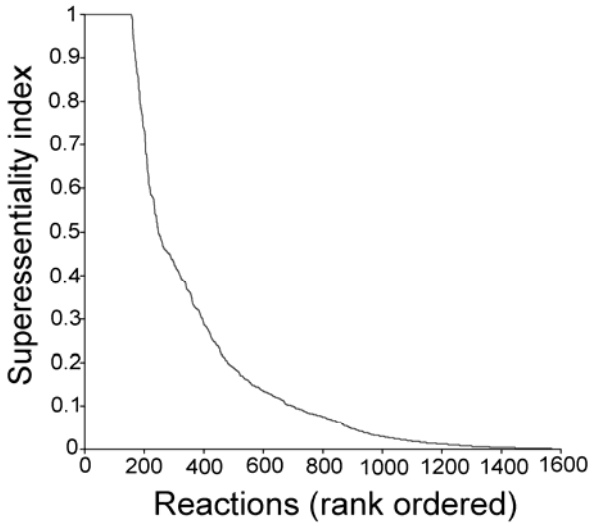
Purine and Pyrimidine Biosynthesis	0.0560	0.0172	4.58E-003	8.51E-003	THE
Nucleotide Salvage Pathway	0.0400	0.0938	1.25E-002	2.72E-002	LTE
Lipopolysaccharide Biosynthesis / Recycling	0.0720	0.0487	7.19E-002	1.87E-001	-
Membrane Lipid Metabolism	0.0080	0.0301	8.50E-002	2.76E-001	-
Threonine and Lysine Metabolism	0.0160	0.0129	2.61E-001	1.13E+000	-
Murein Biosynthesis	0.0080	0.0100	3.59E-001	3.59E-001	-
Methionine Metabolism	0.0080	0.0100	3.59E-001	4.67E+000	-

Significant *p*-values are highlighted in red (Benjamini-Hochberg test p-value < 0.05) [80]. Abbreviations: HTE – higher than expected; LTE – Lower than expected.

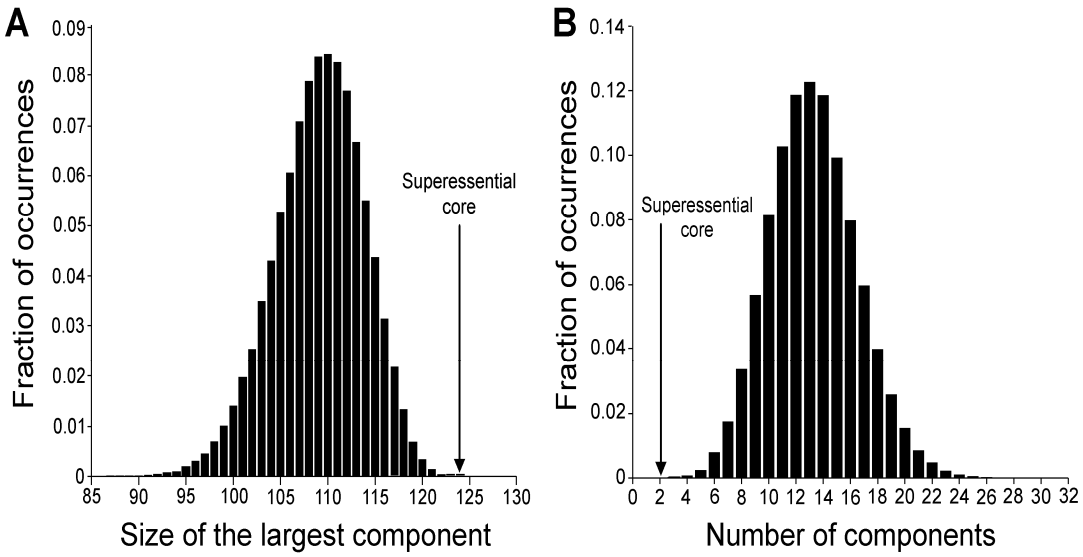
Supplementary table S4 – Genome occurrence for all reaction-coding genes of essential reactions for growth on complex *in vivo* environments. The table also shows the *p*-value for the association between genome occurrence and superessentiality.

Environment	Number of absolutely superessential reactions	Number of reactions whose enzyme-coding genes have an $I_{GO} > 0.5$	*p*-value (number of reactions with $I_{GO} > 0.5$ cannot be explained by chance alone)
Salmonella typhimirium LT2	100	81	10^{-5}, n= 10^5
Mycobacterium tuberculosis H37Rv	108	86	10^{-5}, n= 10^5
Pseudomonas aeruginosa PAO1	101	87	10^{-5}, n= 10^5
Mycoplasma pneumoniae	97	78	10^{-5}, n= 10^5
Synthetic complete medium	99	85	10^{-5}, n= 10^5

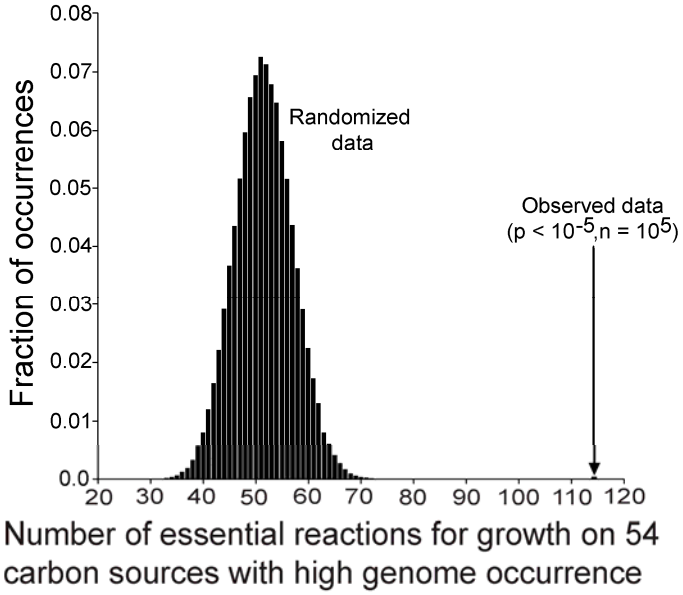
Supplementary figures



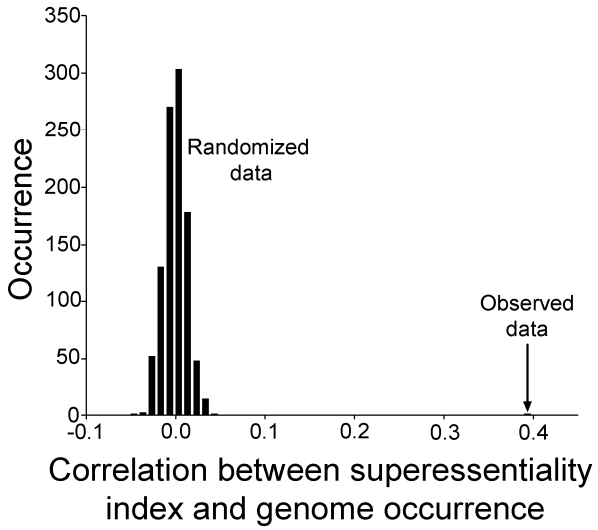
Supplementary figure S1: Rank histogram of the superessentiality index of 1569 reactions. These reactions are essential for growth on 54 carbon sources in 500 random viable metabolic networks. The plateau to the left of the plot corresponds to 155 absolutely superessential reactions ($I_{SE} = 1$).



Supplementary figure S2: The superessential core forms a highly compact subset of metabolism. (A) The figure shows the distribution of the size of the largest strongly connected component in 10^5 random sets of 125 reactions. The size of the largest component in the actual superessential core (124) is also shown, indicating its low probability to occur by chance alone ($p < 10^{-5}$). (B) The figure shows the distribution of the number of strongly connected components from 10^5 random sets of 125 reactions. The number of components in the actual superessential core (2 components) is also shown, indicating the low probability that this number of components occurs by chance alone ($p < 10^{-5}$).



Supplementary figure S3: The vertical axis shows the probability of a reaction being absolutely superessential and having a high genome occurrence ($I_{GO} \geq 0.5$). The horizontal axis shows the number of such reactions in each sample. The plot demonstrates that the data contains many more reactions that are absolutely superessential and have a high genome occurrence than randomized data. Data is based on genome occurrence of absolutely superessential reactions in 500 random metabolic networks viable on 54 carbon sources.



Supplementary figure S4 - The figure shows the distribution of the Spearman's index of statistical association between the superessentiality index and genome occurrence for randomized data and observed data. The association in the actual observed data is significantly higher ($p < 10^{-5}$, $n = 10^5$) than in randomized data. All data is based on superessentiality indices and genome occurrences of all reactions in known reaction universe for viability on 54 different carbon sources.

2.7 References

1. Neidhardt, F. & Ingraham, J. *Escherichia coli and Salmonella typhimurium: Cellular and Molecular Biology*. **1,** (American Society for Microbiology, Washington, DC, 1987).

2. Almaas, E., Oltvai, Z. N. & Barabasi, A. L. The activity reaction core and plasticity of metabolic networks. *PLoS Comput Biol* **1,** e68 (2005).

3. Segrè, D., Vitkup, D. & Church, G. M. Analysis of optimality in natural and perturbed metabolic networks. *Proc. Natl. Acad. Sci. U. S. A.* **99,** 15112–7 (2002).

4. Edwards, J. S. & Palsson, B. O. Robustness analysis of the Escherichia coli metabolic network. *Biotechnol Prog* **16,** 927–939 (2000).

5. Price, N. D., Papin, J. A. & Palsson, B. O. Determination of Redundancy and Systems Properties of the Metabolic Network of Helicobacter pylori Using Genome-Scale Extreme Pathway Analysis. *Genome Res.* **12,** 760–769 (2002).

6. Gerdes, S. Y. *et al.* Experimental determination and system level analysis of essential genes in Escherichia coli MG1655. *J. Bacteriol.* **185,** 5673–84 (2003).

7. Pál, C. *et al.* Chance and necessity in the evolution of minimal metabolic networks. *Nature* **440,** 667–70 (2006).

8. Shlomi, T., Berkman, O. & Ruppin, E. Regulatory on/off minimization of metabolic flux changes after genetic perturbations. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 7695–700 (2005).

9. Wilding, E. I. *et al.* Identification, evolution, and essentiality of the mevalonate pathway for isopentenyl diphosphate biosynthesis in gram-positive cocci. *J. Bacteriol.* **182,** 4319–27 (2000).

10. Chaudhuri, R. R. *et al.* Comprehensive identification of essential Staphylococcus aureus genes using Transposon-Mediated Differential Hybridisation (TMDH). *BMC Genomics* **10,** 291 (2009).

11. Baba, T. *et al.* Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol. Syst. Biol.* **2,** 2006.0008 (2006).

12. Feist, A. M. *et al.* A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol. Syst. Biol.* **3,** 121 (2007).

13. Wang, Z. & Zhang, J. Abundant indispensable redundancies in cellular metabolic networks. *Genome Biol Evol* **1,** 23–33 (2009).

14. Papp, B., Pal, C. & Hurst, L. D. Metabolic network analysis of the causes and evolution of enzyme dispensability in yeast. *Nature* **429,** 661–664 (2004).

15. Joyce, A. R. *et al.* Experimental and computational assessment of conditionally essential genes in Escherichia coli. *J Bacteriol* **188,** 8259–8271 (2006).

16. Dowell, R. D. *et al.* Genotype to phenotype: a complex problem. *Science (80-.).* **328,** 469 (2010).

17. Payne, D. J., Gwynn, M. N., Holmes, D. J. & Pompliano, D. L. Drugs for bad bugs: confronting the challenges of antibacterial discovery. *Nat Rev Drug Discov* **6,** 29–40 (2007).

18. Gadad, A. K., Mahajanshetti, C. S., Nimbalkar, S. & Raichurkar, A. Synthesis and antibacterial activity of some 5-guanylhydrazone/thiocyanato-6-arylimidazo[2,1-b]-1,3, 4-thiadiazole-2-sulfonamide derivatives. *Eur. J. Med. Chem.* **35,** 853–7 (2000).

19. Wiesner, J., Borrmann, S. & Jomaa, H. Fosmidomycin for the treatment of malaria. *Parasitol. Res.* **90 Suppl 2,** S71–6 (2003).

20. Banerjee, A. *et al.* inhA, a gene encoding a target for isoniazid and ethionamide in Mycobacterium tuberculosis. *Science* **263,** 227–30 (1994).

21. Timmins, G. S. & Deretic, V. Mechanisms of action of isoniazid. *Mol. Microbiol.* **62,** 1220–7 (2006).

22. Matias Rodrigues, J. F. & Wagner, A. Evolutionary plasticity and innovations in complex metabolic reaction networks. *PLoS Comput. Biol.* **5,** e1000613 (2009).

23. Samal, A., Matias Rodrigues, J. F., Jost, J., Martin, O. C. & Wagner, A. Genotype networks in metabolic reaction spaces. *BMC Syst. Biol.* **4,** 30 (2010).

24. Fong, S. S., Joyce, A. R. & Palsson, B. Ø. Parallel adaptive evolution cultures of Escherichia coli lead to convergent growth phenotypes with different gene expression states. *Genome Res.* **15,** 1365–72 (2005).

25. Fong, S. S., Marciniak, J. Y. & Palsson, B. Ø. O. Description and Interpretation of Adaptive Evolution of Escherichia coli K-12 MG1655 by Using a Genome-Scale In Silico Metabolic Model. *J. Bacteriol.* **185,** 6400–6408 (2003).

26. Ibarra, R. U., Edwards, J. S. & Palsson, B. O. Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. *Nature* **420,** 186–9 (2002).

27. Goto, S., Nishioka, T. & Kanehisa, M. LIGAND: chemical database of enzyme reactions. *Nucleic Acids Res.* **28,** 380–2 (2000).

28. Goto, S., Okuno, Y., Hattori, M., Nishioka, T. & Kanehisa, M. LIGAND: database of chemical compounds and reactions in biological pathways. *Nucleic Acids Res.* **30,** 402–4 (2002).

29. Barve, A., Rodrigues, J. F. M. & Wagner, A. Superessential reactions in metabolic networks. *Proc. Natl. Acad. Sci. U. S. A.* **109,** E1121–30 (2012).

30. Blattner, F. R. *et al.* The complete genome sequence of Escherichia coli K-12. *Science (80-.).* **277,** 1453–62 (1997).

31. Mengin-Lecreulx, D. & van Heijenoort, J. Characterization of the essential gene glmM encoding phosphoglucosamine mutase in Escherichia coli. *J. Biol. Chem.* **271,** 32–9 (1996).

32. Kawai, S., Mori, S., Mukai, T., Hashimoto, W. & Murata, K. Molecular characterization of Escherichia coli NAD kinase. *Eur. J. Biochem.* **268,** 4359–65 (2001).

33. Burgard, A. P., Nikolaev, E. V, Schilling, C. H. & Maranas, C. D. Flux coupling analysis of genome-scale metabolic network reconstructions. *Genome Res.* **14,** 301–12 (2004).

34. Maltsev, N., Glass, E. M., Ovchinnikova, G. & Gu, Z. Molecular mechanisms involved in robustness of yeast central metabolism against null mutations. *J. Biochem.* **137,** 177–87 (2005).

35. Samal, A. *et al.* Low degree metabolites explain essential reactions and enhance modularity in biological networks. *BMC Bioinformatics* **7,** 118 (2006).

36. Vitkup, D., Kharchenko, P. & Wagner, A. Influence of metabolic network structure and function on enzyme evolution. *Genome Biol.* **7,** R39 (2006).

37. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28,** 27–30 (2000).

38. Becker, D. *et al.* Robust Salmonella metabolism limits possibilities for new antimicrobials. *Nature* **440,** 303–7 (2006).

39. Zhang, Y.-M. & Rock, C. O. Thematic review series: Glycerolipids. Acyltransferases in bacterial glycerophospholipid synthesis. *J. Lipid Res.* **49,** 1867–74 (2008).

40. Brand, L. A. & Strauss, E. Characterization of a new pantothenate kinase isoform from Helicobacter pylori. *J. Biol. Chem.* **280,** 20185–8 (2005).

41. Leonardi, R. *et al.* A pantothenate kinase from Staphylococcus aureus refractory to feedback regulation by coenzyme A. *J. Biol. Chem.* **280,** 3314–22 (2005).

42. Koonin, E. V, Mushegian, A. R. & Bork, P. Non-orthologous gene displacement. *Trends Genet.* **12,** 334–6 (1996).

43. Galperin, M. Y., Walker, D. R. & Koonin, E. V. Analogous enzymes: independent inventions in enzyme evolution. *Genome Res.* **8,** 779–90 (1998).

44. Houten, S. M. & Waterham, H. R. Nonorthologous gene displacement of phosphomevalonate kinase. *Mol. Genet. Metab.* **72,** 273–6 (2001).

45. Khersonsky, O. & Tawfik, D. S. Enzyme promiscuity: a mechanistic and evolutionary perspective. *Annu. Rev. Biochem.* **79,** 471–505 (2010).

46. Nam, H. *et al.* Network context and selection in the evolution to enzyme specificity. *Science* **337,** 1101–4 (2012).

47. Raghunathan, A., Reed, J., Shin, S., Palsson, B. & Daefler, S. Constraint-based analysis of metabolic capacity of Salmonella typhimurium during host-pathogen interaction. *BMC Syst. Biol.* **3,** 38 (2009).

48. Fang, X., Wallqvist, A. & Reifman, J. Development and analysis of an in vivo-compatible metabolic network of Mycobacterium tuberculosis. *BMC Syst. Biol.* **4,** 160 (2010).

49. Oberhardt, M. A., Goldberg, J. B., Hogardt, M. & Papin, J. A. Metabolic network analysis of Pseudomonas aeruginosa during chronic cystic fibrosis lung infection. *J. Bacteriol.* **192,** 5534–48 (2010).

50. Yus, E. *et al.* Impact of genome reduction on bacterial metabolism and its regulation. *Science* **326,** 1263–8 (2009).

51. Förster, J., Famili, I., Palsson, B. & Nielsen, J. Large-scale evaluation of in silico gene deletions in Saccharomyces cerevisiae. *Omi. a J. Integr. Biol.* **7,** 193–202 (2003).

52. Becker, S. A. & Palsson, B. Ø. Genome-scale reconstruction of the metabolic network in Staphylococcus aureus N315: an initial draft to the two-dimensional annotation. *BMC Microbiol.* **5,** 8 (2005).

53. Henry, C. S. *et al.* High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nat. Biotechnol.* **28,** 977–82 (2010).

54. Matias Rodrigues, J. F. & Wagner, A. Genotype networks, innovation, and robustness in sulfur metabolism. *BMC Syst Biol* **5,** 39 (2011).

55. Shen, Y. *et al.* Blueprint for antimicrobial hit discovery targeting metabolic networks. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 1082–7 (2010).

56. Knuth, K., Niesalla, H., Hueck, C. J. & Fuchs, T. M. Large-scale identification of essential Salmonella genes by trapping lethal insertions. *Mol. Microbiol.* **51,** 1729–1744 (2004).

57. Sassetti, C. M. & Rubin, E. J. Genetic requirements for mycobacterial survival during infection. *Proc. Natl. Acad. Sci. U. S. A.* **100,** 12989–94 (2003).

58. Sassetti, C. M., Boyd, D. H. & Rubin, E. J. Genes required for mycobacterial growth defined by high density mutagenesis. *Mol. Microbiol.* **48,** 77–84 (2003).

59. Jacobs, M. A. *et al.* Comprehensive transposon mutant library of Pseudomonas aeruginosa. *Proc. Natl. Acad. Sci. U. S. A.* **100,** 14339–44 (2003).

60. Liberati, N. T. *et al.* An ordered, nonredundant library of Pseudomonas aeruginosa strain PA14 transposon insertion mutants. *Proc. Natl. Acad. Sci. U. S. A.* **103,** 2833–8 (2006).

61. Motter, A. E., Gulbahce, N., Almaas, E. & Barabási, A.-L. Predicting synthetic rescues in metabolic networks. *Mol. Syst. Biol.* **4,** 168 (2008).

62. Arora, N., Banerjee, A. K. & Murty, U. S. N. In silico characterization of Shikimate Kinase of Shigella flexneri: A potential drug target. *Interdiscip. Sci.* **2,** 280–90 (2010).

63. Dias, M. V *et al.* Chorismate synthase: an attractive target for drug development against orphan diseases. *Curr Drug Targets* **8,** 437–444 (2007).

64. Pillai, B. *et al.* Structural insights into stereochemical inversion by diaminopimelate epimerase: an antibacterial drug target. *Proc. Natl. Acad. Sci. U. S. A.* **103,** 8668–73 (2006).

65. Pillai, B. *et al.* Dynamics of catalysis revealed from the crystal structures of mutants of diaminopimelate epimerase. *Biochem. Biophys. Res. Commun.* **363,** 547–53 (2007).

66. Usha, V., Dover, L. G., Roper, D. I., Fütterer, K. & Besra, G. S. Structure of the diaminopimelate epimerase DapF from Mycobacterium tuberculosis. *Acta Crystallogr. D. Biol. Crystallogr.* **65,** 383–7 (2009).

67. Brunetti, L., Galeazzi, R., Orena, M. & Bottoni, A. Catalytic mechanism of l,l-diaminopimelic acid with diaminopimelate epimerase by molecular docking simulations. *J. Mol. Graph. Model.* **26,** 1082–90 (2008).

68. Velasco, A. M., Leguina, J. I. & Lazcano, A. Molecular evolution of the lysine biosynthetic pathways. *J. Mol. Evol.* **55,** 445–59 (2002).

69. Chan, J. N. Y., Nislow, C. & Emili, A. Recent advances and method development for drug target identification. *Trends Pharmacol. Sci.* **31,** 82–8 (2010).

70. Gant, T. W., Zhang, S.-D. & Taylor, E. L. Novel genomic methods for drug discovery and mechanism-based toxicological assessment. *Curr. Opin. Drug Discov. Devel.* **12,** 72–80 (2009).

71. Sioud, M. Main approaches to target discovery and validation. *Methods Mol. Biol.* **360,** 1–12 (2007).

72. Oh, Y.-K., Palsson, B. O., Park, S. M., Schilling, C. H. & Mahadevan, R. Genome-scale reconstruction of metabolic network in Bacillus subtilis based on high-throughput phenotyping and gene essentiality data. *J. Biol. Chem.* **282,** 28791–9 (2007).

73. Oberhardt, M. A., Puchałka, J., Fryer, K. E., Martins dos Santos, V. A. P. & Papin, J. A. Genome-scale metabolic network analysis of the opportunistic pathogen Pseudomonas aeruginosa PAO1. *J. Bacteriol.* **190,** 2790–803 (2008).

74. Jamshidi, N. & Palsson, B. Ø. Investigating the metabolic capabilities of Mycobacterium tuberculosis H37Rv using the in silico strain iNJ661 and proposing alternative drug targets. *BMC Syst. Biol.* **1,** 26 (2007).

75. Price, N. D., Reed, J. L. & Palsson, B. Ø. Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nat. Rev. Microbiol.* **2,** 886–97 (2004).

76. Kauffman, K. J., Prakash, P. & Edwards, J. S. Advances in flux balance analysis. *Curr. Opin. Biotechnol.* **14,** 491–6 (2003).

77. Kanehisa, M. *et al.* From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.* **34,** D354–7 (2006).

78. Binder, K. & Heerman, D. W. *Monte Carlo Simulation in Statistical Physics*. (Springer, 2010).

79. Ma, H.-W. & Zeng, A.-P. The connectivity structure, giant strong component and centrality of metabolic networks. *Bioinformatics* **19,** 1423–30 (2003).

80. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Methodol.* **57,** 289–300 (1995).

81. Wagner, A. & Fell, D. A. The small world inside large metabolic networks. *Proc. Biol. Sci.* **268,** 1803–10 (2001).

82. Salzberg, S. L. Genome re-annotation: a wiki solution? *Genome Biol* **8,** 102 (2007).

83. Jackson, E. N. & Yanofsky, C. Localization of two functions of the phosphoribosyl anthranilate transferase of Escherichia coli to distinct regions of the polypeptide chain. *J. Bacteriol.* **117,** 502–8 (1974).

84. Kim, C., Song, S. & Park, C. The D-allose operon of Escherichia coli K-12. *J. Bacteriol.* **179,** 7631–7637 (1997).

85. Chandran, V. & Luisi, B. F. Recognition of enolase in the Escherichia coli RNA degradosome. *J Mol Biol* **358,** 8–15 (2006).

86. Morita, T., Kawamoto, H., Mizota, T., Inada, T. & Aiba, H. Enolase in the RNA degradosome plays a crucial role in the rapid decay of glucose transporter mRNA in the response to phosphosugar stress in Escherichia coli. *Mol Microbiol* **54,** 1063–1075 (2004).

87. Kumar, J. K., Tabor, S. & Richardson, C. C. Proteomic analysis of thioredoxin-targeted proteins in Escherichia coli. *Proc Natl Acad Sci U S A* **101,** 3759–3764 (2004).

88. Leichert, L. I. & Jakob, U. Protein thiol modifications visualized in vivo. *PLoS Biol.* **2,** e333 (2004).

89. Brown, S. A., Palmer, K. L. & Whiteley, M. Revisiting the host as a growth medium. *Nat. Rev. Microbiol.* **6,** 657–66 (2008).

90. Schramek, N. *et al.* Reaction mechanism of GTP cyclohydrolase I: single turnover experiments using a kinetically competent reaction intermediate. *J. Mol. Biol.* **316,** 829–37 (2002).

3. A latent capacity for evolutionary innovation through exaptation in complex metabolic systems

Aditya Barve[1,2] and Andreas Wagner [1,2,3]

[1] Institute of Evolutionary Biology and Environmental Studies, Bldg. Y27, University of Zurich, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland

[2] The Swiss Institute of Bioinformatics, Bioinformatics, Quartier Sorge, Batiment Genopode, 1015 Lausanne, Switzerland.

[3] The Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA

Part of this chapter was published in *Nature*. 500, 203–206 (2013). [doi: 10.1038/nature12301]

3.1 Abstract

Some evolutionary innovations may originate non-adaptively as pre-adaptations or exaptations, which are by-products of other adaptive traits. Examples include feathers, which originated for thermal insulation but later adopted a role in flight, and lens crystallins, light-refracting proteins that originated as enzymes. The incidence of non-adaptive trait origins has profound implications for evolutionary biology, but it has thus far not been possible to study this incidence systematically. We here study it in metabolism, one of the most ancient biological systems that is central to all life. We analyse metabolic traits of great adaptive importance, the ability of a metabolic reaction network to synthesize all biomass from a single (sole) source of carbon and energy. We take advantage of novel computational methods to randomly sample many metabolic networks that can sustain life on any given carbon source, but that contain an otherwise random set of known biochemical reactions. We show that such random networks, required to be viable on one carbon source C, are typically also viable on multiple other carbon sources C_{new} that were not targets of selection. For example, viability on glucose may entail viability on up to 44 other sole carbon sources. These carbon sources need not be biochemically related, and their identity varies widely among different metabolic networks. Thus, any one adaptation in these metabolic systems typically entails multiple potential exaptations. Metabolic systems thus contain a latent potential for evolutionary innovations with non-adaptive origins. Our observations suggest that many more metabolic traits than currently appreciated may have non-adaptive origins. They also challenge our ability to distinguish adaptive from non-adaptive traits.

3.2 Introduction

How evolutionary adaptations and innovations originate is one of the most profound questions in evolutionary biology. Previous work [1,2] emphasizes the importance of exaptations, also sometimes called pre-adaptations, for this origination. These are traits whose benefits to an organism are unrelated to the reasons for their origination; they are features that originally serve one (or no) function, and become later co-opted for a different purpose [1–5]. Examples of exaptations range from macroscopic to the molecular [6], and abound also in human evolution [7], no number of examples could reveal how important exaptations are in the origination of adaptations in general. This limitation of case studies can be overcome in those biological systems where it is possible to study systematically many genotypes and the phenotypes they form [8–12].

One of these systems is metabolism. The metabolic genotype of an organism encodes a metabolic reaction network with hundreds of enzyme-catalysed chemical reactions. One of metabolism's fundamental tasks is to synthesize small biomass precursor molecules from environmental molecules, such as different organic carbon sources. An organism or metabolic network is *viable* on a carbon source, if it is able to synthesize all biomass molecules from this source. Anecdotal evidence shows that this ability can sometimes originate as a pre-adaptation [13,14]. For example, laboratory evolution of *Pseudomonas putida* for increased biomass yield on xylose as a carbon source produces strains that utilize arabinose as efficiently as xylose, even though the ancestral strains did not utilize arabinose [14]. Thus, viability on arabinose can be a by-product of increased viability on xylose. We here carry out a more systematic analysis of whether such exaptations are typical or unusual in metabolic systems.

Our analysis relies on the ability to predict a metabolic phenotype from a metabolic genotype with the constraint-based method of flux balance analysis (see methods, section 3.5), to study not just one metabolic network but to explore systematically a vast space of possible metabolic networks. The members of this space can be described as follows. The currently known 'universe' of biochemical reactions comprises more than 5,000 chemical reactions with well-defined substrates and products. In the metabolic network of any one organism, however, only a fraction of

these reactions take place, enabling us to describe this network through a binary presence/absence pattern of enzyme-catalysed reactions in the known reaction universe. Recent methods based on Markov chain Monte Carlo (MCMC) sampling (see methods, section 3.5) allow a systematic exploration of this space; that is, they permit the creation of arbitrarily large and uniform samples of networks with a given phenotype [12]. This sampling is based on long random walks through metabolic network space, where each step in a walk adds or eliminates a metabolic reaction from a metabolic network, with the only constraint being that the network remains viable on a focal carbon source. The starting point of the MCMC random walk is the *Escherichia coli* metabolic network, which we know a priori to be viable on different carbon sources [15]. Here we use this approach to create random samples of metabolic networks that are viable on a given set of carbon sources. We refer to such networks as random viable networks.

In our analysis, we focus on 50 different carbon sources (supplementary table S1, section 3.6), many of which are common, and because they are biologically relevant[21]. For each carbon source *C*, we create a sample of 500 random viable networks that are viable on *C*, if *C* is provided as the sole carbon source. We then use FBA to determine the viability of these networks on each of the 49 other carbon sources. This approach allows us to ask whether viability on carbon source *C* usually entail viability on other carbon sources. If so, on how many? How does their number and identity depend on *C*? Do they share biochemical features? How different are they from one another? The answers to these and related questions show that exaptations are ubiquitous in metabolism.

3.3 Results

Networks required to be viable on glucose are also viable on multiple other carbon sources.

We began our analysis with a sample of 500 random networks that were viable on glucose as the sole carbon source (see methods, section 2.5). Each network can synthesize 63 essential biomass precursors in an aerobic minimal environment containing glucose as the only carbon source. These 63 molecules are the biomass precursors of *E. coli*. We used them because they are well-characterized, and many of

their molecules are important in most free-living organisms [15,16]. Importantly, we did not require that these 500 networks are viable on any carbon source except glucose.

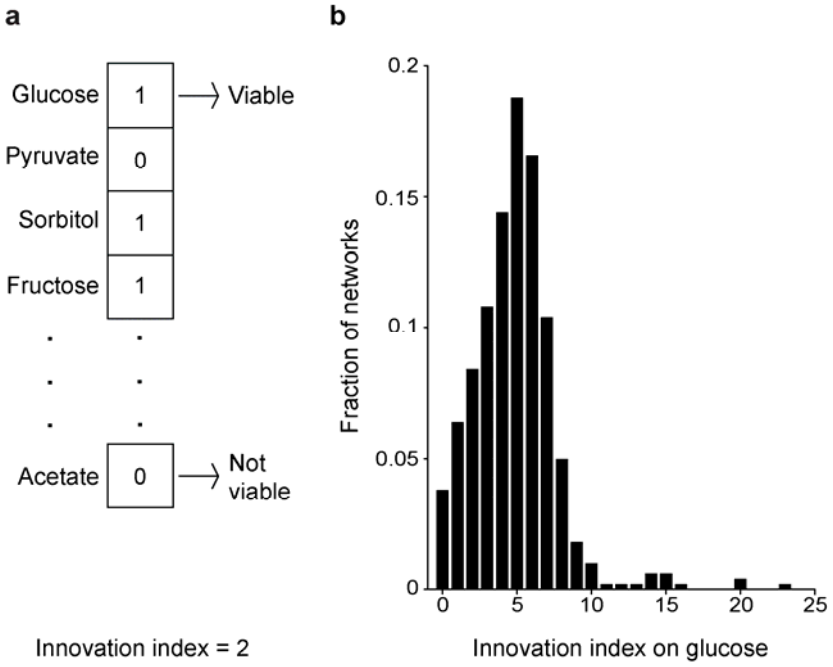


Figure 3.1: (a) The binary innovation vector of a hypothetical metabolic network that is viable on glucose. The vector shows that the random network is viable on glucose, sorbitol and fructose (marked by 1), but not viable on pyruvate and acetate (marked by 0). The innovation index of this network ($I_{Glucose} = 2$) denotes the number of additional carbon sources the network is viable on. (b) The distribution of innovation indices for 500 random networks viable on glucose. Only 4 percent of networks have $I_{Glucose} = 0$, meaning that they are viable only on glucose.

We first examined whether these networks were viable on each of the 49 other carbon sources. The information resulting from this analysis can be represented, for each network, as a binary 'innovation vector' whose *i*-th entry equals one if the network is viable on carbon source C_i, and otherwise zero (figure 3.1a). We define the *innovation index* $I_{Glucose}$ of a network as the number of additional carbon sources that each network is viable on. The distribution of this index is shown in figure 3.1b. Fully 96 percent of networks are viable on other carbon sources in addition to glucose ($I > 0$). The mean innovation index is $I = 4.86$ (standard deviation (s.dev.) = 2.83 carbon sources). This means that networks viable on glucose typically are also viable on

almost 5 additional carbon sources. 18.8 percent of networks (94 networks) are viable on exactly 5 new carbon sources, and 37.4 percent (187) of networks are viable on 6 or more carbon sources. Viability on each such carbon source is a potential exaptation. It is a mere by-product of viability on glucose, and can become a key adaptation whenever this carbon source is the sole carbon source.

In the supplementary results (section 3.6), we show that most of the additional carbon sources to which a metabolic network is pre-adapted differ between different random viable networks (supplementary results, supplementary figures S1 and S2, section 3.6). We show that changing only connected reactions to generate random metabolic networks (see methods, section 3.5) further increases exaptation (supplementary results, supplementary figure S3, section 3.6). We also show that more complex metabolic networks, networks with more reactions, have greater potential for exaptation (supplementary results and supplementary figure S4, section 3.6). Finally, we quantify the potential for exaptation in our sample of 500 metabolic networks selected for growth on glucose (supplementary results, section 3.6) and show that the majority of the carbon sources confer viability on at least one network in our sample.

Glucose is not atypical in its ability to facilitate pre-adaptation.

We next asked whether the ability to grow on multiple additional carbon sources is a peculiarity of networks viable on glucose. To this end, we sampled, for each of our remaining 49 carbon sources, 500 random metabolic networks viable on this carbon source (for a total of 49 x 500 = 24500 sampled networks).

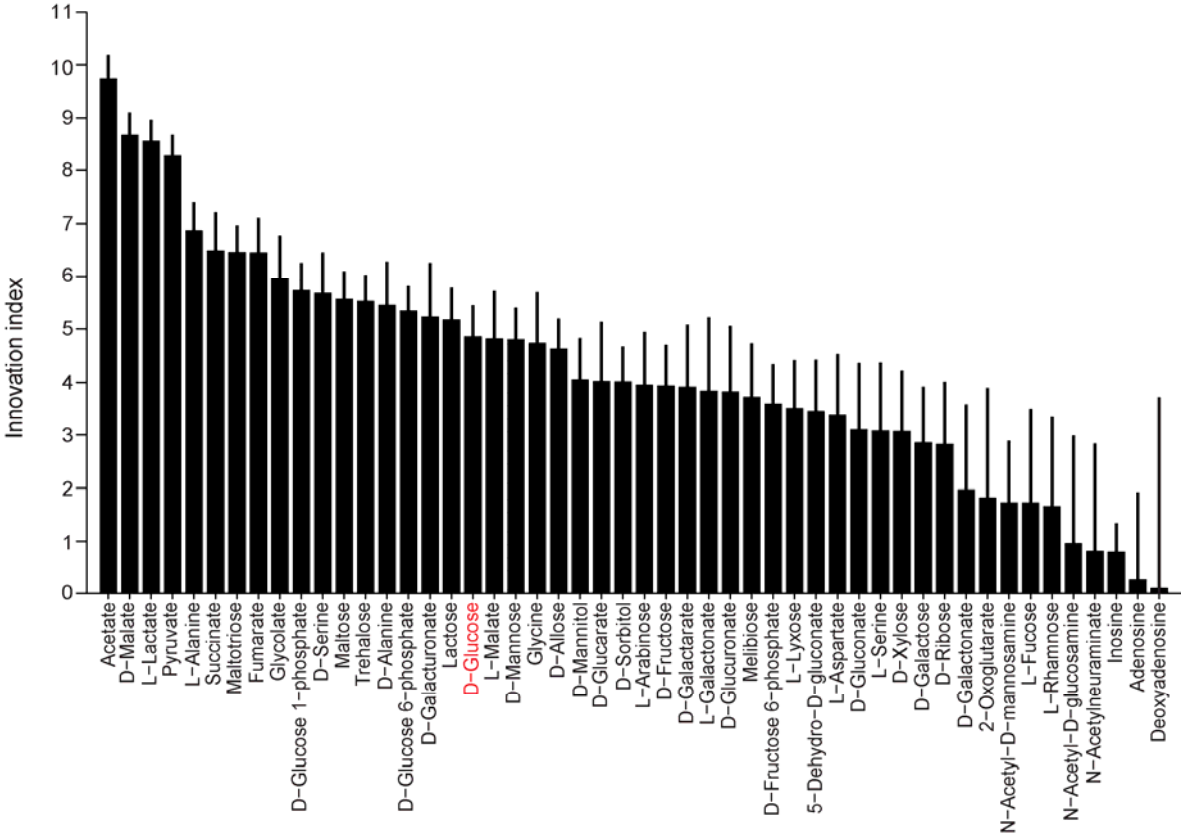


Figure 3.2: For each of 50 carbon sources *C* (horizontal axis), the figure indicates the mean innovation index (bar) and its coefficient of variation (lines) for 500 random networks required to be viable on carbon source *C*. Note the broad distribution of the index. Some carbon sources such as acetate allow viability on more than nine additional carbon sources on average, while others, such as deoxyadenosine support viability on fewer than one additional carbon sources. The innovation index of glucose (in red) is typical compared to other carbon sources.

We then computed the distribution of the innovation index I_C for each carbon source *C*. Figure 3.2 shows the mean of this distribution (bars) and its coefficient of variation (vertical lines), that is, the ratio of the standard deviation to the mean. The figure shows that glucose (highlighted in red) is by no means unusual. 36 percent (18) carbon sources have an even greater average innovation index than glucose. For example, acetate allows viability on the greatest number (9.75) of additional carbon sources. Conversely, some carbon sources such as adenosine ($I_{Adenosine}$=0.27) and deoxyadenosine ($I_{Deoxyadenosine}$ = 0.1) allow growth on fewer additional carbon sources than glucose. Carbon sources with a small average innovation index– they entail viability on few additional carbon sources – are also more variable in this innovation

index (supplementary figure S5, Spearman's $\rho = -0.82$, $p < 10^{-101}$). In the supplementary material, we show that even though any one carbon source may confer growth on only few additional carbon sources in any one network (figure 3.2), when considering all networks in a sample, it may still allow pre-adaptation to most other carbon sources (supplementary results, supplementary figure S6, section 3.6).

In sum, what we observed for glucose is not unusual but typical. Viability on any one carbon source *C* usually entails viability on multiple other carbon sources, whose number and identity can vary with *C*. Viability on never before encountered carbon sources is thus a typical metabolic property.

Metabolically close carbon sources show the highest potential for pre-adaptation. The centre path of figure 3.3a shows a hypothetical metabolic pathway that leads from a carbon source *C* to a source C_{new} (boxed area) and from there through (possibly multiple) further metabolic reactions to the synthesis of biomass. Figure 3.3b shows the same scenario, except that C and C_{new} are separated by several further reactions. It is possible that random networks viable on *C* are more likely to be viable also on C_{new}, if C_{new} is closer to *C*, i.e., if they are separated by fewer metabolic reactions, as in the scenario of figure 3.3a. The reason is that in this case, metabolite C_{new} may be less easy to by-pass through an alternative pathway that originates somewhere between *C* and C_{new} (left-most sequence of arrows in figure 3.3b).

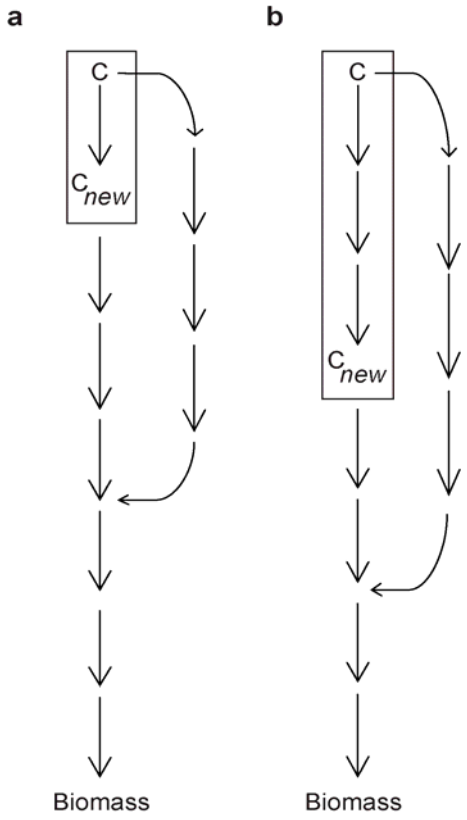


Figure 3.3: (a) The left side of the figure shows a hypothetical carbon source C_{new}, which can be synthesized from some carbon source C in one reaction (arrow), and which leads through multiple further reaction to the synthesis of biomass. Some metabolic networks may have an alternative metabolic pathway that by-passes C_{new} altogether (right sequence of arrows). (b) C_{new} and C are separated by multiple reactions. The fewer reactions separate C and C_{new}, the more likely it may be that both are involved in the synthesis of biomass, that they are not by-passed by some alternative metabolic pathway, and that therefore viability on C implies viability on C_{new}.

To test this hypothesis, we computed metabolic distance as the minimal number of metabolic reactions separating pairs of carbon sources (C, C_{new}), for all possible pairs that can be formed from our 50 carbon sources (see supplementary methods). The maximal distance is six reactions. We then analysed networks selected to be viable on the carbon source glucose. We divided the 49 carbon sources C_{new} different from glucose into two categories, those on which more than the median number of networks in a sample are viable (see distribution in supplementary figure S2 in section 3.6), and those on which fewer than this median number are viable. The average metabolic distance of carbon sources to glucose in the two categories is 2.26 (s.dev. = 0.82) and 3.6 (s.dev. = 1.14), respectively, a difference that is statistically significant (Mann-Whitney U-test, $p = 0.007$). This means that carbon sources C_{new} with a greater incidence of pre-adaptation are metabolically closer to glucose.

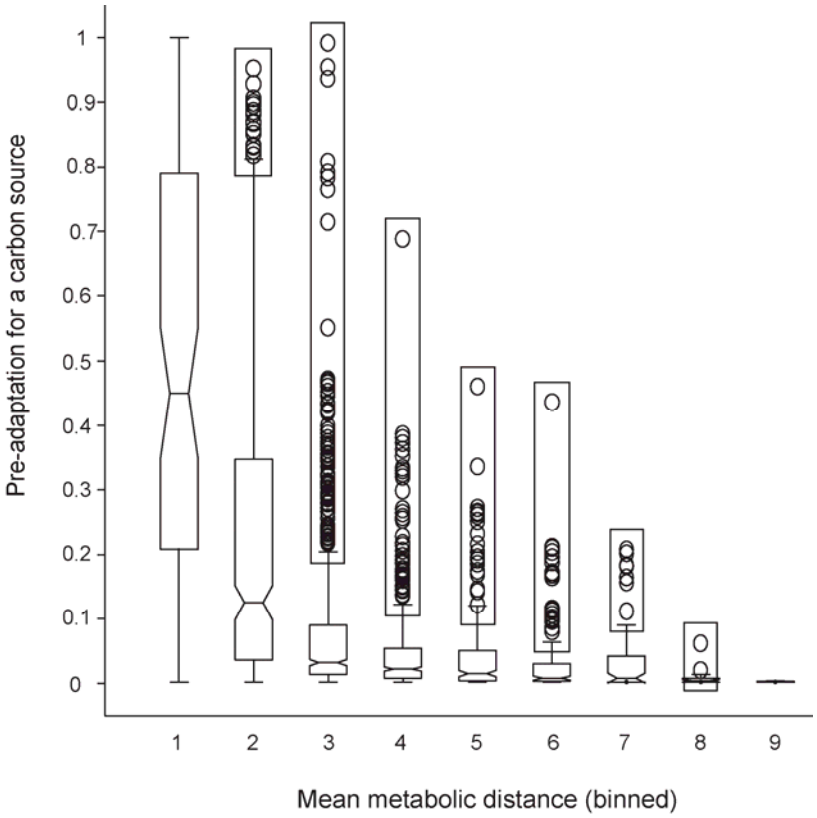


Figure 3.4: The horizontal axis indicates the mean number of reactions that separate *C* and C_{new} in networks that are viable on both *C* and C_{new}, binned into integer intervals corresponding to the floor of their numerical value. The vertical axis indicates the fraction of random metabolic networks required to be viable on carbon source *C* that are additionally viable on C_{new}. Note that the potential for innovation decreases with increasing distance. The central mark in each box denotes the median fraction of networks, the edges indicate the 25th and the 75th percentiles, and whiskers indicate ± 2.7σ and 99.3 coverage and open circles indicate outliers.

In a second, more extensive analysis, we studied the 50 samples of 500 random metabolic networks where networks in each sample were required to be viable on a different one of our 50 carbon sources *C*. For each sample (carbon source *C*), and for each of the other 49 possible carbon sources C_{new}, we asked whether the metabolic distance between *C* and C_{new} is correlated with the fraction of networks that are also viable on C_{new}. To do this, we used metabolic networks that were selected for growth on *C* and additionally viable on C_{new} (supplementary methods). We then computed the mean metabolic distance and binned the distances. The results, pooled for all networks are shown on the vertical axis of figure 3.4, whose horizontal axis reflects the mean metabolic distance (binned into 9 bins). If a carbon source C_{new} is closer to a carbon source *C*, then significantly more networks viable on *C* are also be viable on C_{new} (Spearman's $\rho = -0.42$, $p = 10^{-87}$, $n = 1990$). However, the figure also shows that the association is highly noisy, and especially so at low metabolic distances. In the supplementary results, we discuss the possible reasons behind this noisy association

between metabolic distance and exaptation for a carbon source. This analysis was carried out without taking into account reaction irreversibility, but relaxing that assumption yields the same result (Spearman's $\rho = -0.39$, $p = 10^{-57}$, $n = 1601$, see methods in section 3.5). In the supplementary results, we also show that computing distances between pairs of carbon sources C and C_{new} in the universe of reactions does not change our results.

While metabolic 'nearness' cannot explain exaptations involving two carbon sources, we show that biochemical similarities help us understand why a network viable on C might be viable on one additional carbon source C_{n1}, but not on another source C_{n2}. Indeed, exaptations often involve carbon sources with broadly defined biochemical similarities (supplementary results, supplementary figures S7 and S8, section 3.6). For example, glycolytic carbon sources are more likely to entail exaptations for growth on other glycolytic carbon sources, and likewise for gluconeogenic carbon sources, as well as for carbon sources involved in nucleotide metabolism. Furthermore, we also show that pre-adaptation is synergistic when viability is required on two carbon sources C_1 and C_2. That is, the innovation index for a pair of carbon sources is greater than the sum of the innovation indices I_{C1} and I_{C2} (supplementary figure S9, section 3.6). Moreover, such a capacity for pre-adaptation is higher if carbon sources C_1 and C_2 belong to two different clusters (as shown in supplementary figure S8, section 3.6). We also show that environmental generalists capable of surviving on multiple carbon sources may be viable on many more carbon sources than occur in their environment (supplementary results, supplementary tables S2 and S3, supplementary figure S10, section 3.6). Moreover, networks that convert carbon sources efficiently into biomass tend to be pre-adapted to more carbon sources than inefficient networks (supplementary results, supplementary table S4 and supplementary figure S11, section 3.6).

3.4 Discussion

We next discuss a few limitations of our study. First, our analysis is based on current knowledge about the reaction universe. Future work may increase the number of known reactions, but this would not diminish, but could only enhance the spectrum of

possible exaptations. The reason is that additional reactions would allow the utilization of additional carbon sources by some metabolic networks. Second, most of our analysis focused on random networks that are viable on a specific carbon source, and that change through processes akin to horizontal gene transfer and gene deletion. In the wild, selection can affect more than viability, and networks can change through other processes, such as gene duplication. Such factors may affect the incidence of exaptations. An especially important example involves selection favouring networks with a high rate of biomass synthesis, which is necessary to create organisms with short generation times [17]. This particular selective constraint would not affect our conclusions, because we found that networks with high biomass synthesis rates have even greater potential for metabolic innovation than merely viable networks (supplementary figure S7). Third, we considered all necessary nutrient transporters to be present (see methods, section 3.5). If this is not the case, the incidence of exaptation may be reduced. In this regard, we note that 84 percent of *E. coli* transporters can transport multiple molecules [18], and that their substrate specificity can change rapidly [19], thus ameliorating this constraint. Fourth, real metabolic networks may contain larger number of reactions connected to the rest of metabolism than our randomly sampled networks. However, when restricting our analysis to networks in which all reactions are connected, we found an even greater incidence of exaptation than in random networks (see methods in section 3.5 and supplementary results, supplementary figure S3 in section 3.6). Thus, our results provide a lower bound on the incidence of exaptations. Fifth, for reasons of computational feasibility [20], we considered only a limited number of 50 carbon sources. Analysis of more carbon sources could only augment the potential for exaptation we observed. For example, viability on any one carbon source *C* could entail viability on even more additional carbon sources, had we considered more than 50 carbon sources. Finally, most of our analysis is based on sampling a limited number of 500 networks viable on each carbon source, but sampling of 5000 random networks for select carbon sources yielded identical results (supplementary results and supplementary figure S12, section 3.6).

Our observations show that latent metabolic abilities are pervasive features of carbon metabolism. They expose non-adaptive origins of potentially useful carbon source

utilization traits as a universal and inevitable feature of metabolism. However, the ability to form multiple phenotypes is not restricted to metabolic networks. Many metabolic enzymes are promiscuous in nature and capable of utilizing various substrates [21,22], which can further increase network complexity and the potential for exaptation. The ability to form multiple phenotypes also occurs in regulatory circuits[23], which can form different molecular activity patterns, as well as RNA molecules [24], which can form multiple conformations with different biological functions. Systematic analyses of genotype-phenotype relationships are becoming increasingly possible in such systems [25,26], and already hint at exaptive origins of molecular traits. If confirmed in systematic analyses like ours, the pervasiveness of non-adaptive traits may require a re-thinking of the early origins of beneficial traits, before they become subject to preservation by natural selection.

3.5 Methods

Flux balance analysis (FBA)

FBA is a constraint-based computational method [20,27] used to predict synthetic abilities and other properties of large metabolic networks, which are complex systems of enzyme-catalysed chemical reactions. FBA requires information about the stoichiometry of each molecular species participating in the chemical reactions of a metabolic network. This stoichiometric information is represented as a stoichiometric matrix **S** of dimensions $m \times n$, where m denotes the number of metabolites and n denotes the number of reactions in a network [20,27]. FBA also assumes that the network is in a metabolic steady-state, such as would be attained by an exponentially growing microbial population in an unchanging environment. This assumption allows one to impose the constraint of mass conservation on the metabolites in the network. This constraint can be expressed as

Flux balance analysis (FBA) is a constraint-based computational method [20,27] used to predict synthetic abilities and other properties of large metabolic networks, which are complex systems of enzyme-catalysed chemical reactions. FBA requires information about the stoichiometry of each molecular species participating in the chemical

reactions of a metabolic network. This stoichiometric information is represented as a stoichiometric matrix, *S*, of dimensions $m \times n$, where *m* denotes the number of metabolites and *n* denotes the number of reactions in a network [20,27]. FBA also assumes that the network is in a metabolic steady state, such as would be attained by an exponentially growing microbial population in an unchanging environment. This assumption makes it possible to impose the constraint of mass conservation on the metabolites in the network. This constraint can be expressed as $Sv = 0$, where *v* denotes a vector of metabolic fluxes whose entries, v_i, describe the rate at which reaction *i* proceeds. The solutions, or 'allowable' fluxes, of this equation form a large solution space, but not all of these solutions may be of biological interest. To restrict this space to fluxes of interest, FBA uses linear programming to maximize a biologically relevant quantity in the form of a linear objective function $\boldsymbol{Z}$ (ref [27]). Specifically, the linear programming formulation of an FBA problem can be expressed as

$$\max \mathbf{Z} = \max \{\mathbf{c}^T\mathbf{v} \mid \mathbf{Sv} = 0, \mathbf{a} \leq \mathbf{v} \leq \mathbf{b}\}$$

and the individual entries of vectors *a* and *b* respectively contain the minimal and maximal possible fluxes for each reaction in *v*; that is, each entry v_i is bounded from below by a_i and bounded from above by b_i.

We are here interested in predicting if a metabolic network can sustain life in a given spectrum of environments, that is, whether it can synthesize all necessary small biomass molecules (biomass precursors) required for survival and growth. In a free-living bacterium such as *E. coli*, there are more than 60 such molecules, which include 20 proteinaceous amino acids, DNA and RNA nucleotide precursors, lipids, and cofactors. We use the *E. coli* biomass composition [15] to define the objective function and the vector **c**, because most molecules in *E. coli's* biomass would be typically found in free-living organisms. We used the package CLP (1.4, Coin-OR; https://projects/coin-or.org/Clp) to solve the linear programming problems mentioned above.

Chemical environments

Along with the biomass composition and stoichiometric information about a metabolic network, one needs to define one or more chemical environments that contain the nutrients needed to synthesize biomass precursors. We here consider only minimal aerobic growth environments composed of a sole carbon source, along with oxygen, ammonium, inorganic phosphate, sulfate, sodium, potassium, cobalt, iron (Fe^{2+} and Fe^{3+}), protons, water, molybdate, copper, calcium, chloride, magnesium, manganese and zinc [15]. When studying viability of a metabolic network in different environments, we vary the carbon source while keeping all other nutrients constant. When we say, for example, that a particular network is viable on 20 carbon sources, we mean that the network can synthesize all biomass precursors when each of these carbon sources is provided as the *sole* carbon source in a minimal medium. For reasons of computational feasibility, we restrict ourselves to 50 carbon sources (supplementary table 1 in section 3.6). They are all carbon sources on which *E. coli* is known to be viable from experiments [15]. We chose these carbon sources because many of them are prominent, and because they are of known biological relevance, but we emphasize that our observations do not otherwise make a statement about the metabolism of *E. coli* or its close relatives. They apply to metabolic networks that vary much more broadly in reaction composition than any relative of *E. coli,* because of our network sampling approach described below, which effectively randomizes the reaction composition of a microbial metabolism.

The known reaction universe

The known reaction universe is a list of metabolic reactions known to occur in some organisms. For the construction of this universe, we used data from the LIGAND database [28,29] of the Kyoto Encyclopaedia of Genes and Genomes [30,31]. The LIGAND database is divided into two subsets – the REACTION and the COMPOUND database. These two databases together provide information about metabolic reactions, participating chemical compounds, and associated stoichiometric information in an interlinked manner.

As we also described earlier [12,32,33], we specifically used the REACTION and the COMPOUND databases to construct our universe of reactions while excluding - (i) all reactions involving polymer metabolites of unspecified numbers of monomers, or general polymerization reactions with uncertain stoichiometry, (ii) reactions involving glycans, due to their complex structure, (iii) reactions with unbalanced stoichiometry, and (iv), reactions involving complex metabolites without chemical information [31]. The published *E. coli* metabolic model (*i*AF1260) consists of 1397 non-transport reactions [15]. We merged all reactions in the *E. coli* model with the reactions in the KEGG dataset, and retained only the non-duplicate reactions. After these procedures of pruning and merging, our universe of reactions consisted of 5906 non-transport reactions and 5030 metabolites.

Sampling of random viable metabolic networks

In an organism, a metabolic network can change through mutations. They can lead to addition of new reactions, by way of horizontal-gene transfer, or through the evolution of enzymes with novel activities. They can also lead to loss of reactions through loss-of-function mutations in enzyme-coding genes. Natural selection can preserve those changed metabolic networks that are viable in a particular environment. Together, mutational processes and selection may change a metabolic network drastically on long evolutionary time-scale. Recent work has shown that even metabolic networks that differ greatly in their sets of reactions can have the same metabolic phenotype, that is, the same biosynthetic ability [34]. We here employ a recently developed Markov Chain Monte Carlo (MCMC) random sampling [12,32,33,35,36] procedure to generate metabolic networks that are viable in specific environments, but that contain an otherwise random complement of metabolic reactions. Briefly, this procedure involves random walks in the space of all possible networks. During any one such random walk, a metabolic network can change through the addition and deletion of reactions. Although this process resembles the biological evolution of metabolic networks through horizontal gene transfer and (recombination-driven) gene deletions, we here use it for the sole purpose to create random samples of metabolic networks from the space of all such networks [12,36].

In any one MCMC random walk, we keep the total number of reactions at the same number as the starting *E. coli* network (1397; ref [15]), in order to avoid artefacts due to varying reaction network size [12]. Specifically, each mutation step in a random walk involves an addition of a randomly chosen reaction from the reaction universe, followed by a deletion of a randomly chosen metabolic reaction from the metabolic network. We call such a sequence of reaction addition and deletion a reaction swap. Reaction addition does not abolish the viability of a network in any environment. However, reaction deletion might. Thus, after a reaction deletion, we use FBA to ask whether the network is still viable – it can synthesize all biomass precursors -- in the specified environment. If so, we accept the deletion; otherwise, we reject it and choose another reaction for deletion at random, until we have found a deletion that retains viability. After that, we accept the reaction swap, thus completing a single step in the random walk. We do not subject transport reactions to reaction swaps. These reactions are therefore present in all networks generated by our random walk.

Any MCMC random walk begins from a single starting network, in our case that of *E. coli*. The theory behind MCMC sampling [12,36] , shows that it is important to carry out as many reaction swaps as possible for MCMC to 'erase' the random walker's similarity ('memory') of the initial network. The reason is that successive genotypes in a random walk are strongly correlated in their properties, because they differ by only one reaction pair. These correlations fade with an increasing number of reaction swaps. Because we are interested in analyzing growth phenotypes of networks, correlations to the initial network would result in identification of growth on carbon sources similar to those of the starting network. In past work [12,32] , we found that for the network sizes that we use (1397 reactions), 3×10^3 reactions swaps are sufficient to erase the similarity of the final network to the starting network. To err on the side of caution, we thus carried out 5×10^3 reaction swaps before beginning to sample, and sample a network every 5×10^3 reaction swaps thereafter. In this way, we generated samples of 500 random viable metabolic networks through an MCMC random walk of 2.5×10^6 reaction swaps. We carried out different random walks to sample networks viable on different carbon sources.

For some of our analyses, we also sampled random metabolic networks of sizes different from that of the *E. coli* metabolic network. To do this, we followed a previously established procedure [12,32,33] to create a starting network for an MCMC random walk that has the desired size. This procedure first converts the known universe of reactions into a 'global' metabolic network by including the *E. coli* transport reactions in it. Not surprisingly, this global network can produce all biomass components and is therefore viable on all carbon sources studied here. We used this global network to successively delete a sequence of randomly chosen reactions in the following way. After each reaction deletion, FBA is used to ask whether the network is still viable on a given carbon source. If so, the deletion is accepted; otherwise another reaction is chosen at random for deletion. We deleted in this way as many reactions as needed to generate a network of the desired size. We then used this network as the starting network for an MCMC random walk, as described above, to generate samples of 500 random viable networks.

Identification of disconnected non-functional reactions and the connected reaction universe

We performed some of our analysis with a version of the reaction universe that does not contain disconnected reactions. Reactions that are not connected to the rest of a metabolic network would be nonfunctional, because they cannot carry a non-zero steady-state metabolic flux, and thus could not contribute to the synthesis of biomass. The genes encoding them would eventually be lost from a genome. (We note that this loss could still take tens of thousands of years, given known deleterious mutation rates and generation times [37,38] , enough for some for other genetic or environmental changes to render these reactions functional.) We define a disconnected reaction as a reaction that does not share any one substrate or any one product with any other reaction in the known reaction universe. We focus here on reactions in the universe rather than in one metabolic network, because an individual network can gain additional reactions that may connect previously disconnected reactions. We note that even this "universal" definition of disconnectedness depends on our current knowledge of biochemistry, as well as on the environment, for the right environment could supply metabolites that connect previously disconnected reactions or pathways

to the rest of a metabolic network. To identify the connected universe, we removed disconnected reactions. Because this removal may render other reactions disconnected, we repeated this process iteratively until no further reactions in the universe became disconnected. In this way, we found 3646 reactions of the 5906 reactions in the universe of reactions to be connected. We used this connected universe in some analyses to generate network samples using the MCMC approach.

Estimation of the metabolic distance between carbon sources

To compute the metabolic distance between a pair of carbon sources *C* and C_{new}, we used the 500 networks selected for growth on a specific carbon source *C*. We first represented a network as a *substrate graph* [39]. In this graph, vertices correspond to metabolites. Two metabolites (vertices) are linked by an edge if the metabolites participate in the same metabolic reaction, be it as an educt or as a product. We excluded 'currency' metabolites from this substrate graph, which are metabolites that transfer small chemical groups and are involved in many reactions [40]. Specifically, we excluded protons, H_2O, ATP (adenosine triphosphate), ADP (adenosine diphosphate), AMP (adenosine monophosphate), NADP(H) (nicotinamide adenosine dinucleotide diphosphate), NAD(H) (nicotinamide adenosine dinucleotide), and Pi (inorganic phosphate), CoA (coenzyme A), hydrogen peroxide, ammonia, ammonium, bicarbonate, GTP (guanosine triphosphate), GDP (guanosine diphosphate), and PPi (diphosphate) that occurred in both the cytoplasmic and periplasmic compartments [15]. In addition we also excluded oxidized and reduced forms of cofactors such as quinone, ubiquinone, glutathione, thioredoxin, flavodoxin and flavin moninucleotide. That is, we eliminated all vertices corresponding to these metabolites when constructing the substrate graph. For each metabolic network, we constructed two substrate graphs, first one wherein the reaction irreversibility was ignored and all reactions were considered reversible, and the second graph wherein irreversibility was taken into account. For a network selected for growth on carbon source *C*, we calculated the shortest distance of *C* to each exapted carbon source C_{new} in the substrate graph of that network, as computed by a breadth-first search [41]. We preformed this analysis for each network in our ensemble of 500 networks viable on a carbon source *C*. The distance between carbon sources *C* and C_{new} was then computed

as a mean of the metabolic distances based on networks viable on both carbon sources.

We also computed metabolic distance for any two carbon sources by representing the universe of reactions as a graph in the above manner. We again constructed two substrate graphs, first one wherein the reaction irreversibility was ignored and all reactions were considered reversible, and the second graph wherein irreversibility was taken into account. Taking irreversibility into account increases the maximal distance to infinity as some carbon sources are connected by irreversible reactions.

Clustering of carbon sources based on the innovation matrix

The entries of the innovation matrix $\boldsymbol{I} = (I_{ij})$ represent the fraction of random metabolic networks that we required to be viable on carbon source C_i, and that were additionally viable on carbon source C_j. To cluster the entries of this matrix, we first computed for all pairs of rows in this matrix the quantity $d = 1-\rho$, where ρ is the Spearman rank correlation coefficient between the row entries. This yielded a new, distance matrix which describes the distances between all pairs of rows. We clustered the rows of $\boldsymbol{I}$ by applying UPGMA (Unweighted Pair Group Method with Arithmetic means [42]), a hierarchical clustering method, to the distance matrix.

Hierarchical clustering with UPGMA classifies data such that the average distance between elements belonging to the same cluster is lower than the average distance between elements belonging to different clusters [12]. UPGMA identified two clusters of glycolytic and gluconeogenic carbon sources, and we wanted to know whether the distances between them were significantly different. To this end, we first calculated the distribution of distances $d = 1-\rho$ for all pairs of row vectors of $\boldsymbol{I}$ *within* each of the two clusters. We called the resulting distance distribution the 'within-cluster' distance distribution. Similarly, we computed the distances *between* any pair of row vectors belonging to two different clusters. These formed a 'between-cluster' distance distribution. We then used the non-parametric Mann-Whitney U-test to check if these two distributions were significantly different.

Estimation of carbon waste production

FBA determines the maximal biomass yield achievable by a network for a given carbon source [27]. However, even when a network produces the maximally achievable yield, not all of the carbon input into the network may be converted into biomass. The non-converted carbon input constitutes carbon waste. Such non-utilized carbon can be secreted in the form of one or more metabolites. For example, in a glucose minimal environment, *E. coli* secretes carbon dioxide and acetate into the extracellular compartment as carbon waste. FBA estimates the amount of each metabolite secreted per unit time [15,27]. To estimate the amount of carbon waste that a random network viable on glucose produces, we first identified the different metabolites that it secretes as waste and then computed the amount of carbon waste per metabolite as the product of carbon atoms in that metabolite and the amount of the metabolite secreted (millimoles per gram dry weight per hour). The total carbon waste produced by a network is computed as the sum of the above quantity for each of the secreted carbon-containing molecules. We repeated the above procedure for each random network in a sample of 500 random networks viable on glucose. We found a total of 62 metabolites that are secreted as waste metabolites in at least one network of our sample of networks viable on glucose.

We carried out all numerical analyses using MATLAB (Mathworks Inc.)

3.6 Supplementary results

Networks selected for growth on glucose are pre-adapted to different carbon sources.

We have shown that networks required to be viable on glucose are also viable on multiple other carbon sources (figure 3.1b). We next inquired whether the additional carbon sources to which a metabolic network is pre-adapted differ between different random viable networks, or whether they are mostly identical. We found that most of these carbon sources differ among networks. Specifically, 91.84 percent (45) of the additional carbon sources occur in the innovation vectors of fewer than 40 percent of the networks (supplementary figure S1). We also computed the fraction of carbon sources that both networks in a pair are viable on, among all carbon sources that at

least one network in a pair is viable on. This distribution (supplementary figure S2a) has a mean of only 31.8 percent. In other words, for almost 70 percent of carbon sources to which one network is pre-adapted, the other network is not pre-adapted.

We next computed the pairwise distance between the innovation vectors $I_{Glucose}$ for all 500 metabolic networks in our sample. This distance indicates the number of additional carbon sources that one but not the other network in a pair is viable on. Its distribution (supplementary figure S2b) has a mean of 5.22 carbon sources (s.dev. = 3.16). That is, two networks differ on average in their viability on 5 carbon sources. The distribution is right-skewed (supplementary figure S2b) and contains networks that differ in their viability on many carbon sources. The two networks with the maximum distance of 26.53 are viable on 5 and 23 carbon sources in addition to glucose, but only one of the additional carbon sources is shared between them.

Higher reaction connectivity increases exaptation.

The MCMC random walk (see methods, section 3.5) entails the addition of a randomly chosen reaction from the universe of reaction, followed by the deletion of a randomly chosen reaction. The deletion is accepted only if the network continues to be viable on the given carbon source. Thus, viability is the only constraint we enforce while sampling random networks. However, reactions that are not connected to the rest of the metabolism cannot carry a non-zero flux and would therefore be non-functional (although they could become connected after additional reaction changes) [43]. To assess the effect of disconnected reactions on exaptation, we identified disconnected reactions and removed them from our universe of reactions (see methods, section 3.5) to generate a connected universe of reactions. We used the connected universe of reactions to generate 500 random networks viable on glucose via MCMC sampling. The average innovation index of these networks is equal to 10.5 (s.dev. = 6.7 carbon sources), which is higher than for networks generated using the complete universe of reactions (Mann-Whitney U-test, $p = 10^{-64}$).

In a further analysis to understand the role of connected reactions, we modified the MCMC random walk in the following manner. We allowed the addition of a randomly chosen reaction only if all of its substrates participated in at least one other

reaction in our network, thus ensuring that only connected reactions can be added to a network. (We note that this procedure no longer guarantees detailed balance [44] and uniform sampling of the space of viable networks.) We generated 500 metabolic networks viable on one carbon source, for each of 50 carbon sources using the modified random walk (a total of 500x50=25,000 networks). We then computed the innovation index for each network. We found that for networks viable on glucose, the mean innovation index is much higher (I = 20.6, s.dev. = 8.52, supplementary figure S3, glucose highlighted in red) than in our original sample of MCMC-generated networks viable on glucose (I = 4.86, s.dev. = 2.83, figure 3.1b). Supplementary figure S3 shows that this is also true for all other carbon sources. For example, networks viable on acetate are viable on 9.75 other carbon sources in our original sample of networks (figure 3.2); while they are viable on 27.67 new carbon sources when we constrain the random walk to adding only connected reactions. Thus, removal of disconnected reactions leads to more exaptation in metabolic networks, presumably because considering only connected reactions allows more alternate routes towards biomass synthesis. We note that the connectedness of a reaction depends on current knowledge of biochemistry, as well as on the environment, for the right environment could supply metabolites that connect previously disconnected reactions or pathways to the rest of a metabolic network.

Large metabolic networks have higher innovation potential. The size of a metabolic network is the number of reactions participating in the network. In most of our analyses, we focus on random networks with the same size as that of the *E. coli* metabolic network (1397 reactions [15]). These have a mean innovation index of 4.86 for viability on glucose. However, it is possible that this index may depend on the number of reactions in a network. Larger networks might be more likely to metabolize carbon sources in addition to those on which selection acts. To find out whether this is the case, we generated six additional samples of 500 random metabolic networks viable on glucose, but where networks in different samples differed in size. Specifically, network sizes ranged between 400 and 1600 reactions. We calculated the mean innovation index for networks in each sample. Supplementary figure 4 shows that innovation is positively correlated with network size (Spearman's $\rho = 0.6$, $p = 10^{-300}$, $n = 3500$). The horizontal axis of supplementary figure S4 denotes the network

size categories we considered, and the vertical axis indicates the mean innovation index for networks in one sample (error bars correspond to one standard deviation). The figure shows that larger, more complex networks indeed have a higher innovation index. The figure also shows that the variability in the innovation index increases as network size increases. That is, the number of additional carbon sources on which a network is viable becomes increasingly variable as network complexity increases.

The networks used in most of our analyses were generated using the complete universe of reactions, as opposed to the connected universe described above. We wanted to find out whether removal of some reactions from the complete universe affects the correlation between metabolic network size and the innovation index. To this end, we removed disconnected reactions (see methods, section 3.5) from the networks of each sample. We found that removal of disconnected reactions did not change the correlation between network size and innovation (Spearman's $\rho = 0.59$, $p = 10^{-300}$, $n = 3500$) that we had observed earlier for the complete universe (Spearman's $\rho = 0.6$, $p = 10^{-300}$, $n = 3500$).

Networks selected for growth on glucose have a high potential for exaptation. We next asked whether the proportion of the 49 carbon sources that confers viability to at least one network in our sample is small or large. In other words, is the potential for exaptation restricted to a modest percentage of carbon sources? The answer is no. Almost 90 percent (44) of the additional 49 carbon sources confer viability to at least one network in networks selected for growth on glucose. We note that this number might be even higher if computational feasibility had not restricted us to samples of 500 networks. In sum, networks viable on glucose are viable on multiple additional carbon sources. Thus, most of the carbon sources to which different networks are pre-adapted differ between networks.

In the preceding section, we showed that most carbon sources can be subject to pre-adaptation in networks selected to be viable on glucose. The same holds for networks selected to be viable on most other carbon sources (supplementary figure S6). Each vertical bar of supplementary figure S6 shows, for a network sample viable on a specific carbon source (horizontal axis), the number of additional carbon sources on

which at least one network in the sample is viable. For 86 percent (43) of samples, this number is greater than 40, and for 94 percent (47) of samples it is greater than 25, meaning that more than half of the additional carbon sources confer viability to at least one network in the sample. The exceptions are inosine, adenosine, and deoxyadenosine, which can give rise to pre-adaptations on only 4, 3, and 3 other carbon sources, respectively. Even though any one carbon source may confer growth on only few additional carbon sources in any one network (figure 3.2), when considering all networks in a sample, it may still allow pre-adaptation to most other carbon sources. For example, viability on xylose allows viability on only three additional carbon sources on average (figure 3.2). However, in a sample of 500 networks viable on xylose, pre-adaptation occurs for 43 carbon sources (supplementary figure S6). Pre-adaptation or exaptation can thus occur for the vast majority of carbon sources we examined.

Metabolically close carbon sources show the highest potential for pre-adaptation. Figure 3.4 shows a noisy association between the average metabolic distance and pre-adaptation, especially at low metabolic distances. That is, even if a carbon source C_{new} can be produced from C in a single step, the fraction of networks that are viable on C_{new} may range widely from 0.05 to almost one (left-most bin in figure 3.4). For example, 92.8 percent networks viable on acetate are additionally viable on pyruvate as well, whereas only two of 500 networks viable on pyruvate are additionally viable on N-acetylneuraminate, even though both carbon sources are only two reactions away from pyruvate.

It merits explanation why a metabolic network is not always viable on a carbon source C_{new} that can be produced from metabolite C in a single step. For example, the median fraction of networks viable on C_{new} at a distance of one from glucose is only 0.21. The reason is illustrated in the right-most sequence of arrows of figures 3.3a and 3.3b. It may be possible to synthesize biomass from carbon source C such that carbon source C_{new} is completely bypassed. For example, D-glucose-6-phosphate (C_{new}) can be produced from glucose (C) in one step. However, not 100 percent but only 77.2 percent of networks viable on glucose are additionally viable on D-glucose-6-phosphate. The remainder (22.8 percent or 114 networks) can bypass D-glucose-6-

phosphate. These 114 networks metabolize glucose with either of two reactions. The first is catalysed by xylose isomerase (enzyme commission number (EC) 5.3.1.5), which can convert glucose into fructose [15,45]. The second is catalysed by glucose dehydrogenase (EC 1.1.5.2), which can convert glucose into gluconate [15,46]. In sum, the highly reticulate nature of metabolism allows alternative pathways to by-pass carbon sources very closely related to C, and thus limits the potential for pre-adaptation for any one carbon source C_{new} [47].

We asked whether computing distances between C and C_{new} in the universe of reactions changed the correlation between distance and the fraction of networks that are viable on C_{new}. To do this, we represented the universe of reactions as substrate graph (see methods, section 3.5). The correlation changed very little (Spearman's $\rho = -0.47$, $p = 10^{-132}$, $n = 2450$). On taking reaction irreversibility into account, 388 pairs (of 2500) of carbon sources have infinite distance. However, the correlation for exaptation and distances between carbon sources C and C_{new} remains unchanged (Spearman's $\rho = -0.39$, $p = 10^{-77}$, $n = 2062$).

Pre-adaptation involves preferably broadly similar carbon sources. We next asked whether any further indicators of biochemical similarity among carbon sources might help understand why a network viable on C might be viable on one additional carbon source C_{n1}, but not on another source C_{n2}. For example, of the 500 random metabolic networks selected for growth on acetate, 89.6 percent networks are also viable on L-serine, which is at a metabolic distance of two from acetate. In contrast, only 6 percent networks are additionally viable on N-acetylneuraminate, which also has distance two from acetate. Is there a difference between L-serine and N-acetylneuraminate that accounts for these differences?

To help us ask this question systematically, we defined an innovation matrix $\boldsymbol{I}$, whose construction is described in supplementary figure 7a. The entries of this matrix I_{ij} contain the fraction of those random metabolic networks that we required to be viable on carbon source C_i, and that were additionally viable on carbon source C_j. The distance between two rows represents differences in the spectrum of carbon sources to which networks required to be viable on C_i and C_j are pre-adapted. We computed a

distance measure based on the Spearman's rank correlation coefficient (see methods, section 3.5) for all pairs of row vectors, thus arriving at a distance matrix for these vectors. We then used hierarchical clustering to group carbon sources (row vectors) that allow pre-adaptation on similar spectra of carbon sources. The results are three very distinct and clearly separable groups of carbon sources reflected by deep and statistically significant branches in a dendrogram (supplementary figure S8). Specifically, the three groups comprise (i) glycolytic carbon sources, which are mainly sugars and feed into the glycolytic pathway (green), (ii) gluconeogenic carbon sources that feed into lower glycolysis or the tricarboxylic acid cycle (purple), and (iii) nucleotide carbon sources (inosine, deoxyadeosine and adenosine, in black). By far the most prominent groups are the glycolytic and gluconeogenic carbon sources, comprising 47 of our 50 carbon sources. The pairwise within-cluster distances of row vectors are significantly lower than the between-cluster distances (Mann-Whitney U-test, $p = 10^{-159}$) for these two clusters.

Supplementary figure S7b shows a heat-map representation of the innovation matrix, with rows and columns organized such that they reflect the clusters we detected. Carbon sources within a cluster favour the utilization of other carbon sources within a cluster, e.g., networks viable on one glycolytic carbon source tend to be viable on other glycolytic carbon sources as well. To go back to our opening example, viability on acetate, a gluconeogenic carbon source, is more likely to entail viability on another gluconeogenic carbon source, such as L-serine, than on N-acetylneuraminate, a glycolytic carbon source.

Pre-adaptation through required viability on two carbon sources is synergistic. In our analysis thus far, we studied samples of random viable networks that we required to be viable on only one carbon source. However, many organisms have to be viable on more than one carbon source in the wild. This raises the question whether the innate capacity for pre-adaptation increases or decreases as one requires viability on multiple carbon sources. For computational feasibility, we restrict ourselves here to analyses of two carbon sources. Specifically, we chose at random 100 pairs of carbon sources, and generated for each carbon source pair (C_1, C_2) 100 metabolic networks required to be viable on both carbon sources. We then asked whether the average innovation

index for these networks $I_{(C1,C2)}$ was greater or smaller than the sum of the innovation indices I_{C1} and I_{C2}. To this end, we calculated the quantity $I_{(C1,C2)} - I_{C1} - I_{C2}$. This quantity would be equal to zero if pre-adaptation was additive whenever viability was required on two carbon sources (C_1, C_2). Supplementary figure S9 indicates that the distribution of $I_{(C1,C2)}$ is displaced to the right of the origin, and significantly different from zero (One sample t-test, $p = 10^{-10}$, $n = 100$). Specifically, for 77 percent of carbon source pairs, the number of additional carbon sources to which pre-adaptation occurs is greater than the sum of the innovation indices I_{C1} and $I_{C2,}$ and for 23 percent pairs it is less. Thus, viability when required on a pair of carbon sources (C_1, C_2) leads to pre-adaptation on more carbon sources than expected from the two carbon sources C_1 and C_2 separately.

We hypothesized that pre-adaptation on a pair of carbon sources would be higher if carbon sources C_1 and C_2 belonged to two different clusters (supplementary figures S7b and S8), because then each source would facilitate pre-adaptation to other carbon sources in its respective cluster. To test this hypothesis, we computed the innovation index $I_{(C1,C2)}$ separately for two groups of carbon source pairs. In the first group, carbon source C_1 belonged to a different cluster than carbon source C_2. In the second group, C_1 and C_2 belonged to the same cluster. The innovation index $I_{(C1,C2)}$ of carbon source pairs belonging to different clusters (mean $I_{(C1,C2)} = 4.02$) was significantly higher than when C_1 and C_2 belonged to the same clusters (mean $I_{(C1,C2)} = 0.6$; Mann-Whitney U-test, $p = 10^{-8}$). Thus, the capacity of pre-adaptation increases when viability is required on a pair of carbon sources that are biochemically dissimilar.

Environmental generalists may be viable on many more carbon sources than occur in their environments. Environment-generalists such as *E. coli* can sustain life on more than 50 carbon sources [15]. Because viability on one carbon source may entail viability on multiple others, *E. coli* may have experienced selection for viability on substantially fewer than the 50 carbon sources we study. In other words, viability on multiple carbon sources may be an indirect by-product of selection on several other carbon sources. In our next analysis, we asked how many fewer carbon sources are required to allow growth on the majority of the 50 carbon sources studied here. To this end, we first generated a sample of 100 random metabolic networks viable on 10

randomly chosen carbon sources and calculated the average innovation index of these networks. We then repeated this procedure for further samples of 100 networks, requiring viability on an increasing number of carbon sources. Supplementary figure S10 shows the average number of carbon sources on which networks are actually viable (vertical axis, error bars indicate one s.dev.), as a function of the number of carbon sources on which viability is required (horizontal axis). The figure demonstrates that pre-adaptation follows a principle of diminishing returns. Networks need to be, on average, viable on almost 49 random carbon sources to show viability on all 50 carbon sources. We restricted ourselves in all our analyses to 50 carbon sources for reasons of computational feasibility, and note that the usefulness of our analysis is limited by this fact. Specifically, networks required to be viable on 49 carbon sources may be viable on many more than the 50 carbon sources we examined. The impression of diminishing returns may results partly from the upper limit we impose on the number of carbon sources.

In contrast to observations about networks required to be viable on multiple carbon sources, our earlier analysis had shown that viability on pairs of carbon sources (C_1, C_2) can entail pre-adaptation on more carbon sources than expected from each of the two carbon sources, especially if (C_1, C_2) are biochemically dissimilar (supplementary results and supplementary figure 9). By extension, it might be possible to choose a modest number (C_1, …, C_n) of carbon sources, such that viability required on each of these carbon sources entails pre-adaptation on a much larger number of carbon sources, e.g., all or most of the 50 carbon sources we study. To identify such groups of carbon sources we pursued the following, heuristic procedure that involves our innovation matrix $\boldsymbol{I}$. Recall that the entries of this matrix I_{ij} contain the fraction of those random metabolic networks that we required to be viable on carbon source C_i, and that were additionally viable on carbon source C_j. For a pre-specified threshold T, we examined each column C_j of this matrix, to see if the largest entry of the column exceeded T, meaning that a fraction T of networks were also viable on C_j when required to be viable on at least one carbon source C_i. This approach resulted in identifying carbon sources that networks were pre-adapted to (C_{new}) when required to be viable on other carbon sources (C) for a specific threshold T.

We used a threshold $T = 0.75$, meaning that for the sample of networks required to be viable on at least any one carbon source C_i, 75 percent or the majority of its networks were additionally viable on another carbon source C_j. With this approach, we found that requiring viability on 34 specific carbon sources should entail viability on 16 further carbon sources. To validate this hypothesis, we generated 500 random metabolic networks viable on the specific 34 carbon sources (when provided as sole carbon sources). We then computed the mean total number of carbon sources these random networks are viable on. Specifically, we found that random networks were viable on a mean of 49.33 carbon sources (s.dev. = 0.84) when required to be viable on the specific 34 carbon sources ($T = 0.75$, supplementary table S2). This means that selection on these 34 specific carbon sources allows networks to be viable on almost all carbon sources we considered. That is, they are pre-adapted to significantly more carbon sources than 500 random networks viable on a randomly chosen set of 34 carbon sources (mean = 42.23, s.dev. = 1.24) (Mann-Whitney U-test, $p = 10^{-169}$, supplementary table S3). We repeated this procedure with varying thresholds, $T = 0.25$ and $T = 0.5$ and again found pre-adaptation to significantly more carbon sources than 500 random networks viable on a randomly chosen set of carbon sources (supplementary table S3). This analysis shows that metabolic networks are pre-adapted to more carbon sources when required to be viable on a specific set of carbon sources. Thus, environment generalists may have benefitted through similar requirement of viability and growth on a subset of carbon sources, which allowed them to be viable on a repertoire of multiple carbon sources.

Less carbon waste means more pre-adaptation. We next report on an association between a network's biomass yield and its innovation index. We found this association when we divided our original sample of 500 random networks required to be viable on glucose into two groups, according to whether a network's biomass yield lay above or below the mean yield. In this analysis, random metabolic networks with a high biomass yield also showed a significantly higher innovation index (Mann-Whitney U-test, $p = 10^{-5}$). Next we generated 500 random networks with a biomass yield on glucose equal to or exceeding that of the *E. coli* metabolic network [15]. This sample of networks showed a mean innovation index (6.04, s.dev. = 3.7) that was significantly higher (Mann-Whitney U-test, $p = 10^{-8}$) than our original sample of

networks ($I_{Glucose}$ = 4.89, s.dev. = 2.83). Thus, networks with higher biomass yield have a higher innovation index.

A high biomass yield may indicate that a network produces less carbon waste. To find out whether this is the case, we calculated the total carbon waste produced by each network in our original sample of random networks viable on glucose (see methods, section 3.5), and found that the biomass yield is indeed negatively correlated with the amount of carbon waste produced by these networks (Spearman's ρ = -0.89, p = 10^{-168}, n = 500). Furthermore, there is a modest yet significant negative correlation between the amount of carbon waste and the innovation index of networks viable on glucose (Spearman's ρ = -0.28, $p = 10^{-10}$, n = 500).

We speculated that in the networks producing more carbon waste (and having low biomass yield), one or more additional carbon sources are excreted as carbon waste and cannot be fed into biomass production. Such networks would then not be pre-adapted for viability on these carbon sources. We first note that carbon waste can be secreted in the form of various metabolites, such as carbon dioxide, acetate, and fumarate, to name a few. We found a total of 62 carbon-containing metabolites are secreted as waste by at least one network in our sample of 500 networks viable on glucose. For each metabolite, and for each of our two samples (high and low biomass yield), we counted the number of networks in which the metabolite is secreted as waste. Supplementary table S4 shows for each potentially secreted metabolite, the number of high and low yield networks that secrete it. Significantly fewer high-yield networks secreted carbon-containing metabolites, when we considered all these metabolites together as a group (Mann-Whitney U-test, p = 0.0081, n = 62). For 92 percent (57 of 62) metabolites, the number of low yield networks secreting the metabolite is higher than the number of high yield metabolites (supplementary table S4). Furthermore, 12 of these 62 metabolites are used as carbon sources as well (supplementary table S4, shown in red). This observation is particularly important for metabolites that can also serve as carbon sources or are linked to carbons sources. For example, acetate, which is also one of our 50 carbon sources, is secreted as waste by 173 low yield metabolic networks, but only by 122 high yield networks. Other examples of carbon sources that are excreted by a greater number of low-yield

networks include 5-dehydro-D-gluconate, D-gluconate, fumarate, glycolate, succinate, pyruvate.

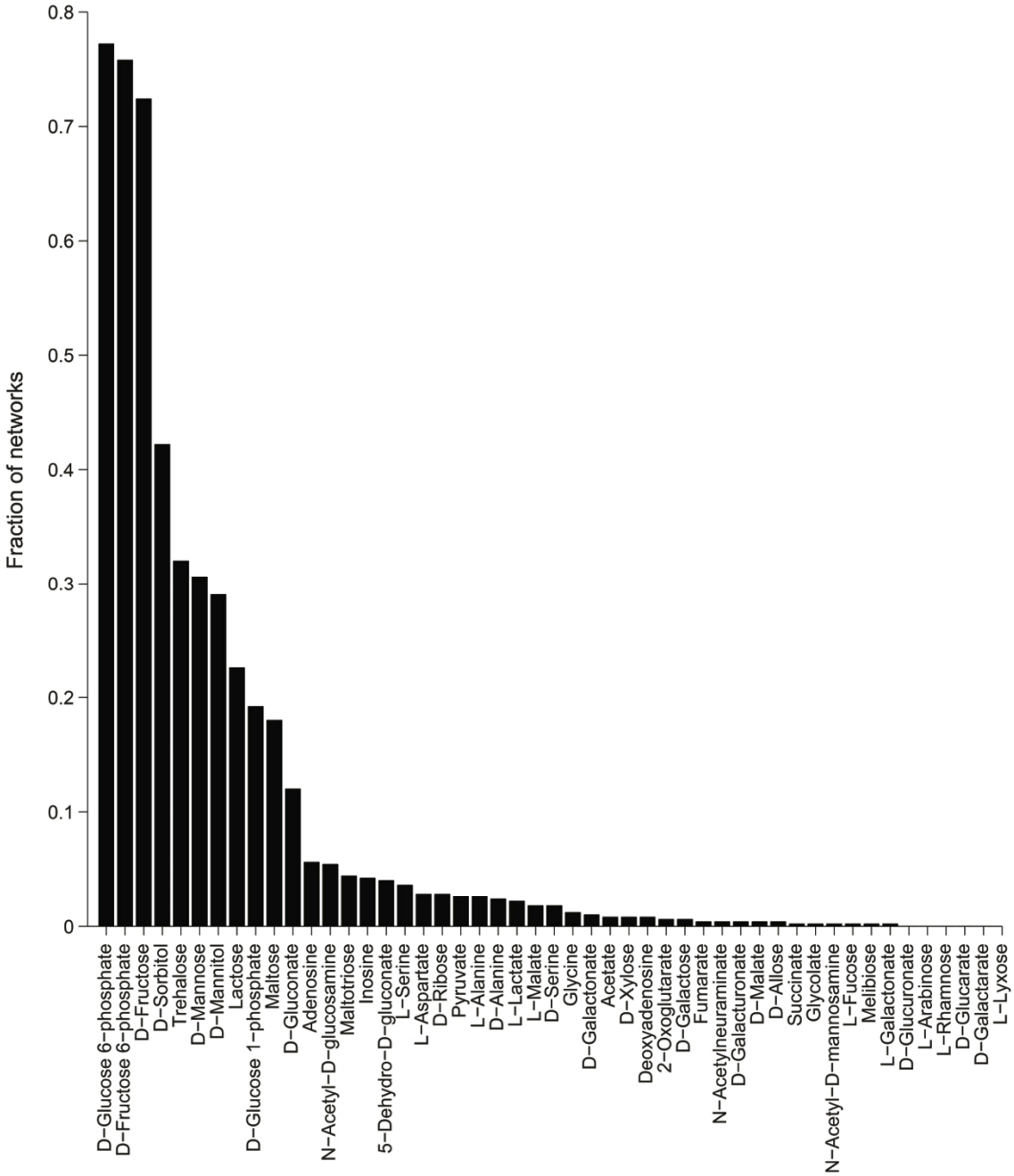
We next asked whether the innovation index correlates with biomass yield per carbon not just for glucose, but for all other carbon sources as well. We define the biomass yield per carbon as the ratio of the biomass yield of a metabolic network to the number of carbons in a particular carbon source. Biomass yield needs to be defined in this manner for this analysis, because different carbon sources contain different numbers of carbon molecules. Supplementary figure S11 shows that the average innovation index on a carbon source C, and the mean biomass yield per carbon show a strong positive correlation (Spearman's $\rho = 0.47$, $p = 0.00057$, $n = 50$). As we mentioned above, a high biomass yield per carbon reflects the efficient conversion of the carbon source into biomass precursors with little waste. Thus, what holds for glucose also holds for other carbon sources.

In sum, a network that converts carbon sources efficiently into biomass tends to have a high innovation index. It tends to be pre-adapted to a larger number of carbon sources. The reason is that the waste products of inefficient metabolic networks include carbon sources. These carbon sources are not utilized by the inefficient network, but can be utilized by an efficient network.

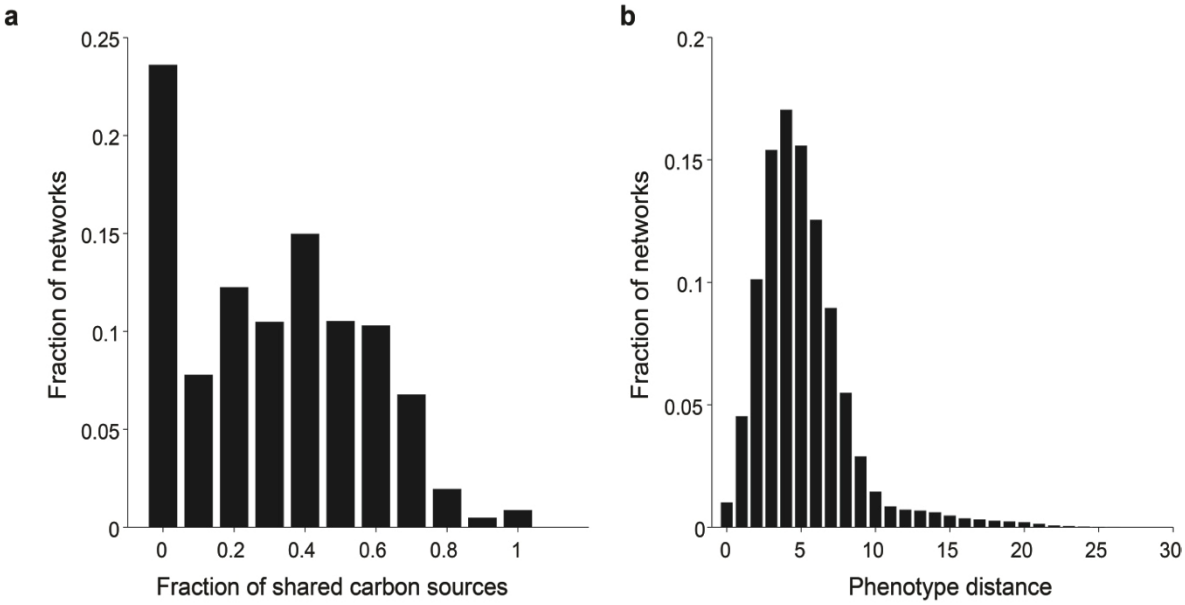
A sample size of 500 networks is sufficient for our analysis. Most of our analysis presented here used 500 random networks viable on a single carbon source. While a sample size of 500 networks has proven to be sufficient for understanding the essentiality of reactions[53], this may not be the case here. To find out, we sampled 5000 networks for each of the following 10 carbon sources: pyruvate, acetate, D-glucose, L-aspartate, L-serine, adenosine, N-acetylneuraminate, trehalose, maltotriose and L-galactonate. These carbon sources have varying innovation indices (figure 3.2), ranging from the highest to the lowest values we observed. We then computed the distribution of the innovation index I_C for each of these ten C carbon sources. Supplementary figure S12 shows the mean of this distribution (bars) and its coefficient of variation (vertical lines), that is, the ratio of the standard deviation to the mean. Black indicates values for the sample of 5000 networks, while grey

indicates values for the original sample of 500 networks. Note that the means are very similar for the two samples of different size. For each carbon source, we also computed the fraction of networks viable on each of the other 49 carbon sources (identical to the innovation matrix explained in the supplementary results, supplementary figure S7a). We then computed the correlation between the entries of this matrix between the resampled and the original ensemble of random networks for each carbon source. The correlation is strong and significant (Spearman's $\rho \geq 0.7$, $p \leq 10^{-7}$, $n = 50$ for all 10 carbon sources. These observations suggest that a sample size of 500 random networks is sufficient for the analyses conducted here.

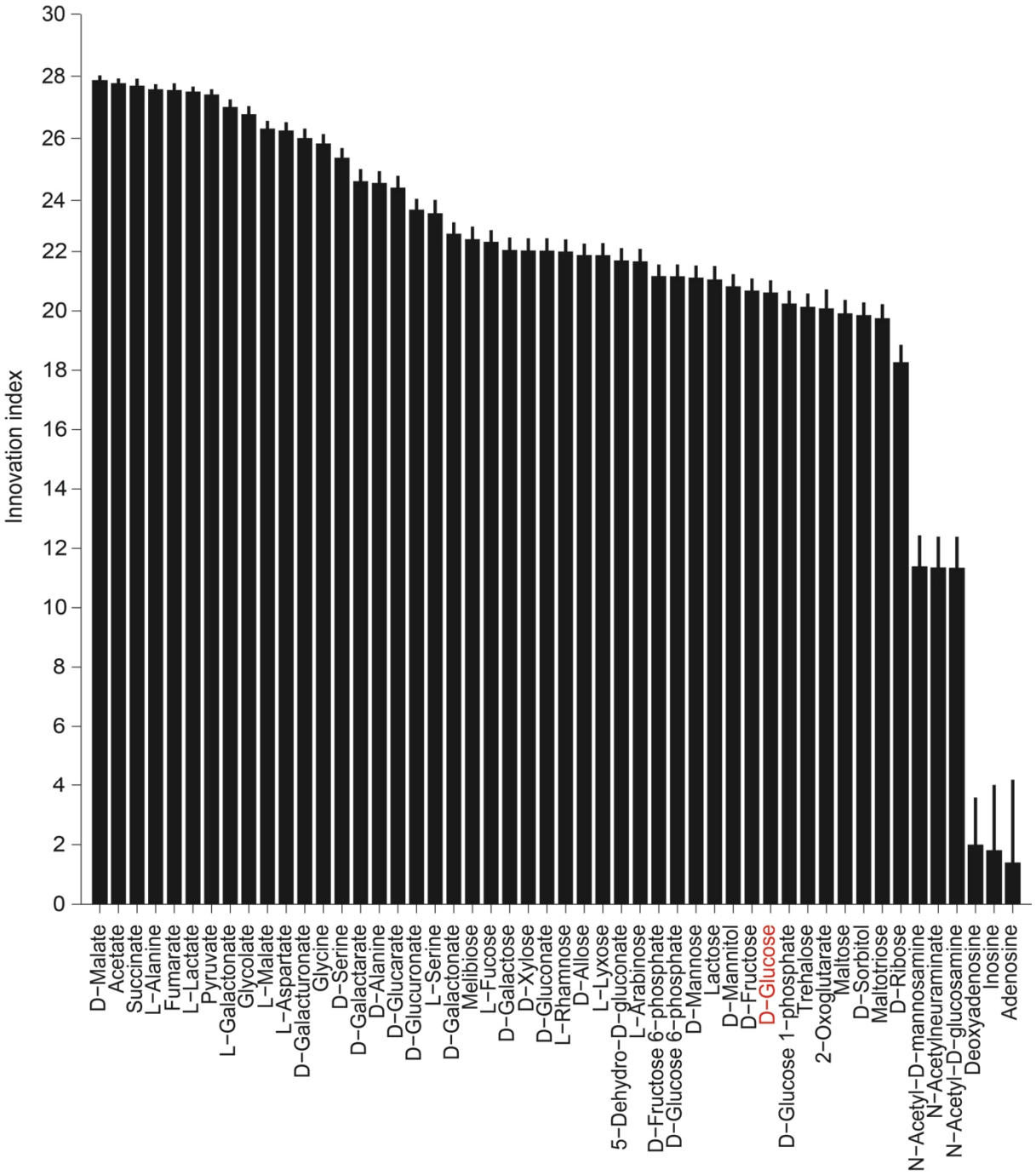
Supplementary figures



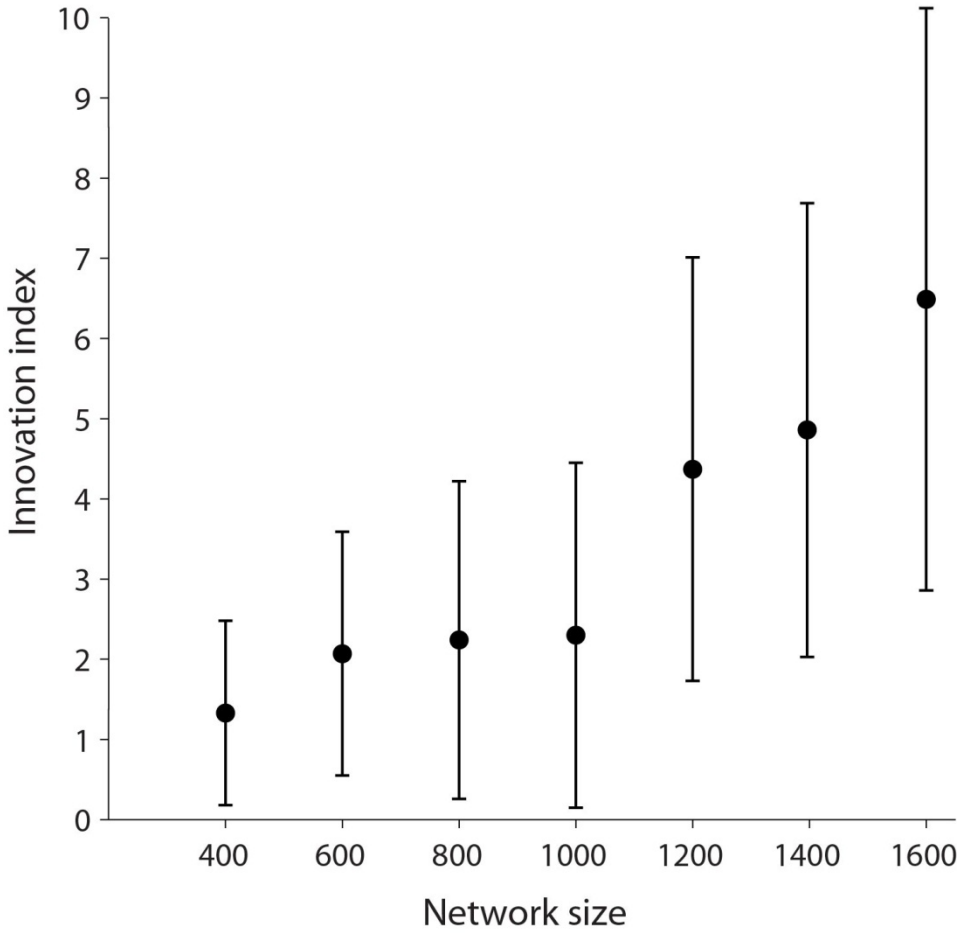
Supplementary figure S1: Different carbon sources differ greatly in their propensity for exaptation. The horizontal axis lists 49 different carbon sources. The vertical axis indicates the fraction of random networks viable on each carbon source (when required to be viable on glucose). Carbon sources are ranked according to the value on the vertical axes. While more than 70 percent of networks are viable on glucose-6-phosphate, fructose-6-phosphate and fructose as additional carbon sources (left-most three bars), most carbon sources allow viability of only a small fraction of sampled networks. Data is based on 500 random networks viable on glucose.



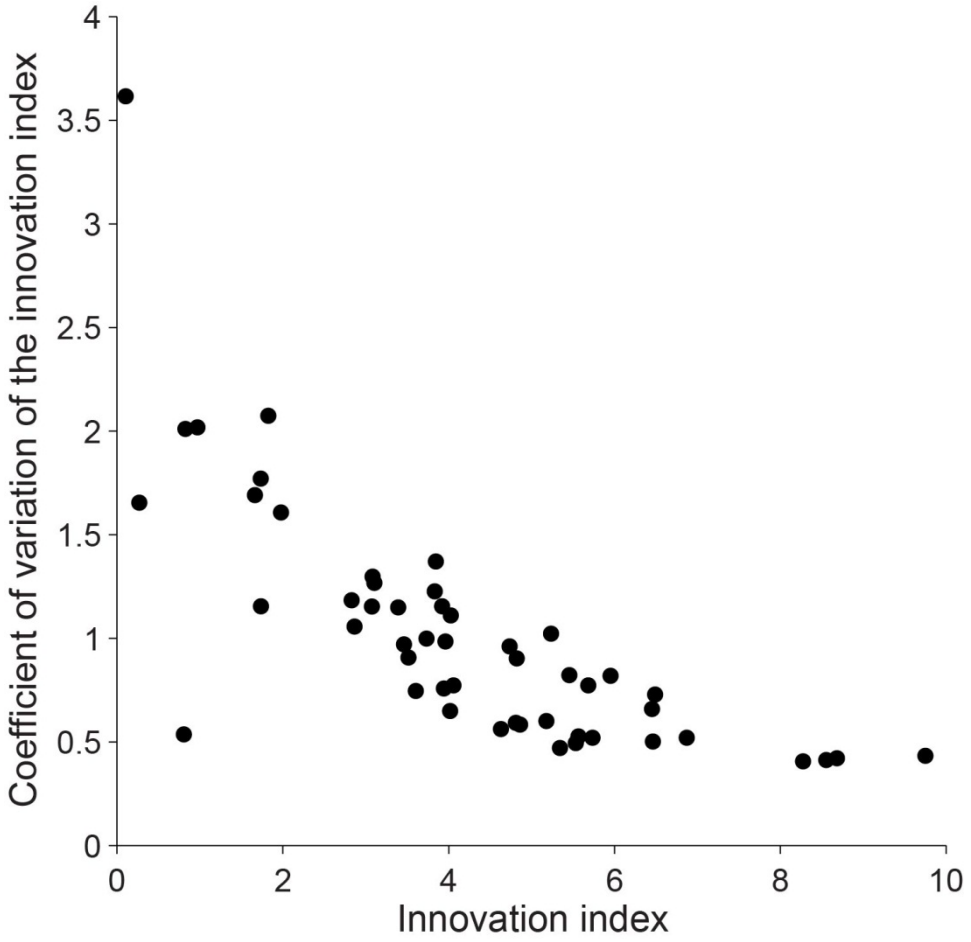
Supplementary figure S2: Majority of the pre-adapted carbon sources differ among networks. (a) The distribution of the number of shared carbon sources in the innovation vector of networks pairs. 23 percent of network pairs do not share any carbon source that they are viable on (except glucose); on average 31.8 percent of carbon sources are shared between a pair of networks. (b) The distribution of the phenotypic distance between network pairs, as computed by the Hamming distance [48] between their innovation vectors. The Hamming distance increases by one for each entry in which two binary vectors differ. The distance thus indicates the number of carbon sources (aside from glucose) that one but not the other network pair is viable on. On average, two networks differ in their viability on 5 carbon sources. Data in (a)-(b) are based on innovation vectors of 500 random networks required to be viable only on glucose.



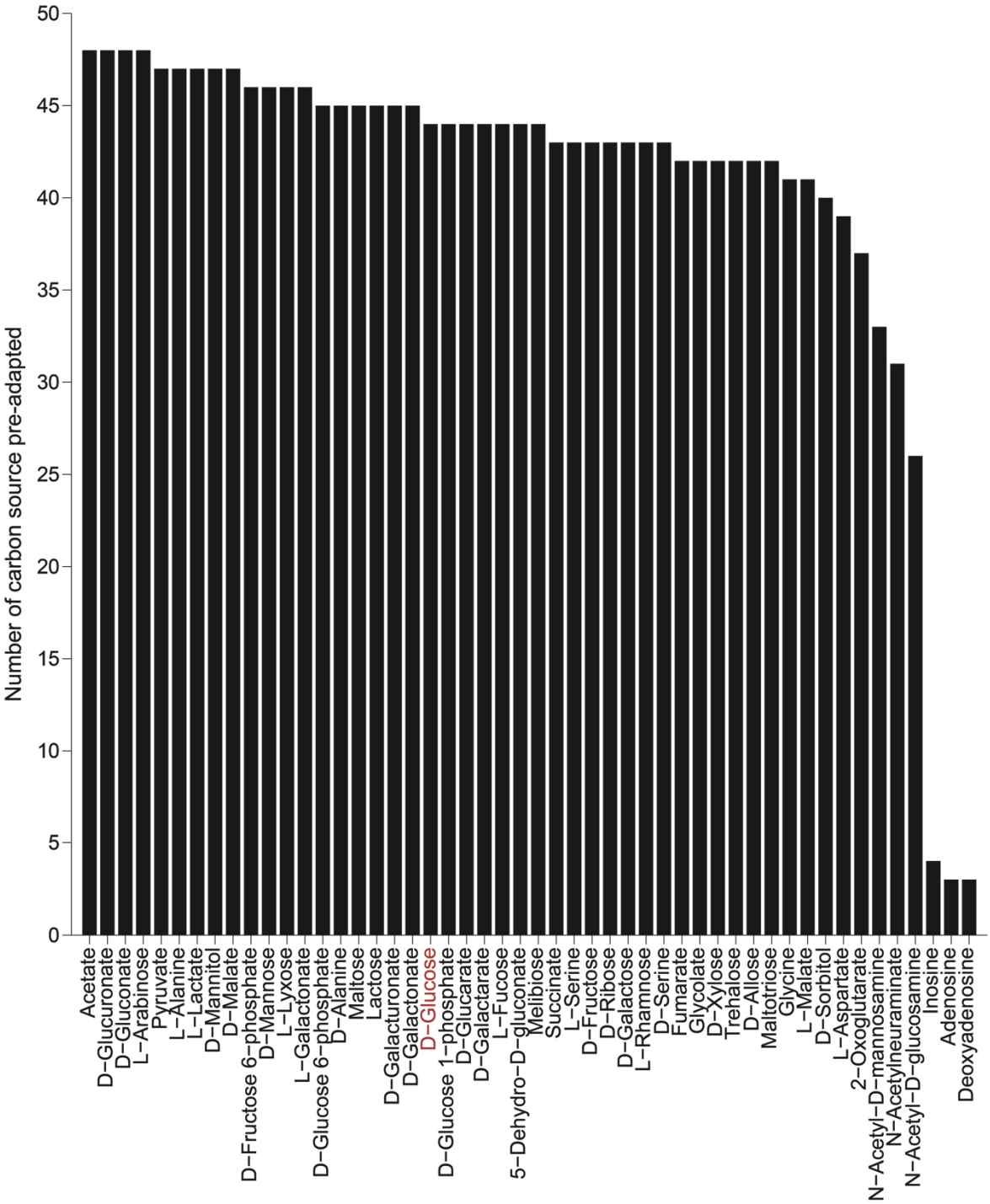
Supplementary figure S3: Innovation potential increases with network connectedness. For each of 50 carbon sources *C* (horizontal axis), the figure indicates the mean innovation index (bar) and its coefficient of variation (lines) for 500 random networks required to be viable on carbon source *C*. Each sample was generated through a sampling process similar to our MCMC sampling, except that we only allowed a reaction to be added to a network, if it is connected to the network through its substrates or products. Considering only such connected networks increases the potential for exaptation. For example, the innovation index of glucose (in red) is much higher than in the original sample of networks (figure 3.2).



Supplementary figure S4: The potential for innovation increases with network complexity. The horizontal axis shows network size (network complexity) in numbers of reactions. A size of 1400 reactions corresponds approximately to the size of the *E. coli* metabolic network (1397 reactions [15]) used in our other analyses. The vertical axis shows the mean innovation index of networks with a given size. Higher network complexity allows viability on a larger number of additional carbon sources. Data for each network size class are based on samples of 500 random networks viable on glucose, i.e. a total of 3500 networks.



Supplementary figure S5: For carbon sources with a high innovation index, this index is less variable. The horizontal axis indicates the mean innovation index, and the vertical axis indicates the coefficient of variation of this index. The data is the same as for figure 3.2, i.e., each data point is based on a sample of 500 random networks required to be viable on one of 50 carbon sources (n = 25000). Note that the coefficient of variation decreases with increasing innovation index.



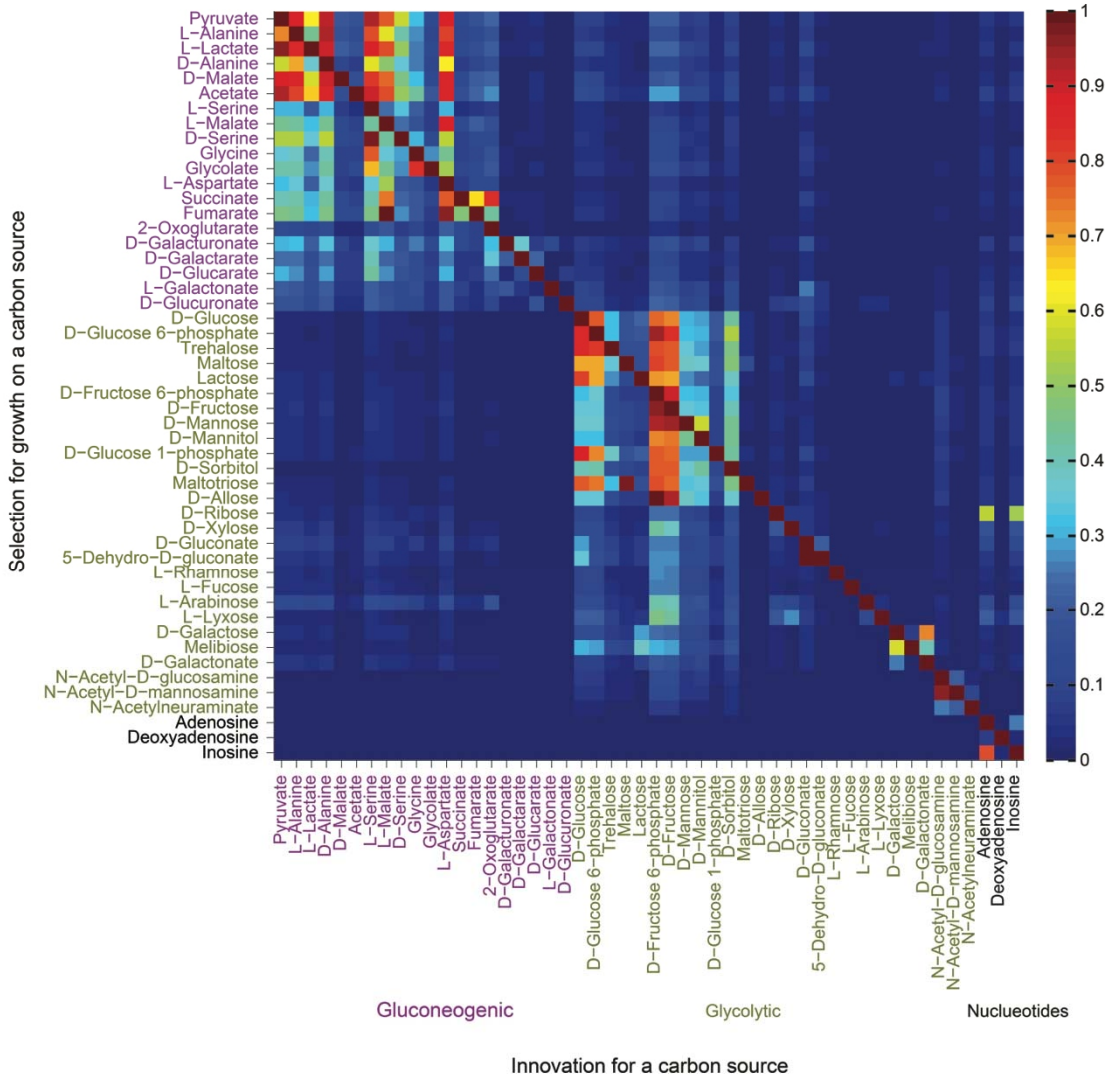
Supplementary figure S6: Pre-adaptation occurs for the vast majority of carbon sources. For each of 50 carbon sources *C* (horizontal axis) the height of the vertical bar above each carbon source indicates the total number of the other 49 carbon sources on which at least one network in the sample is viable. For instance, acetate allows viability (pre-adaptation) on 48 other carbon sources, while deoxyadeosine and adenosine allow viability on only 3 other carbon sources. Data in the figure are based on samples of 500 random viable networks for each carbon source, i.e., on a total of 500x50=25,000 sampled networks.

a

Additional carbon sources C_{new}

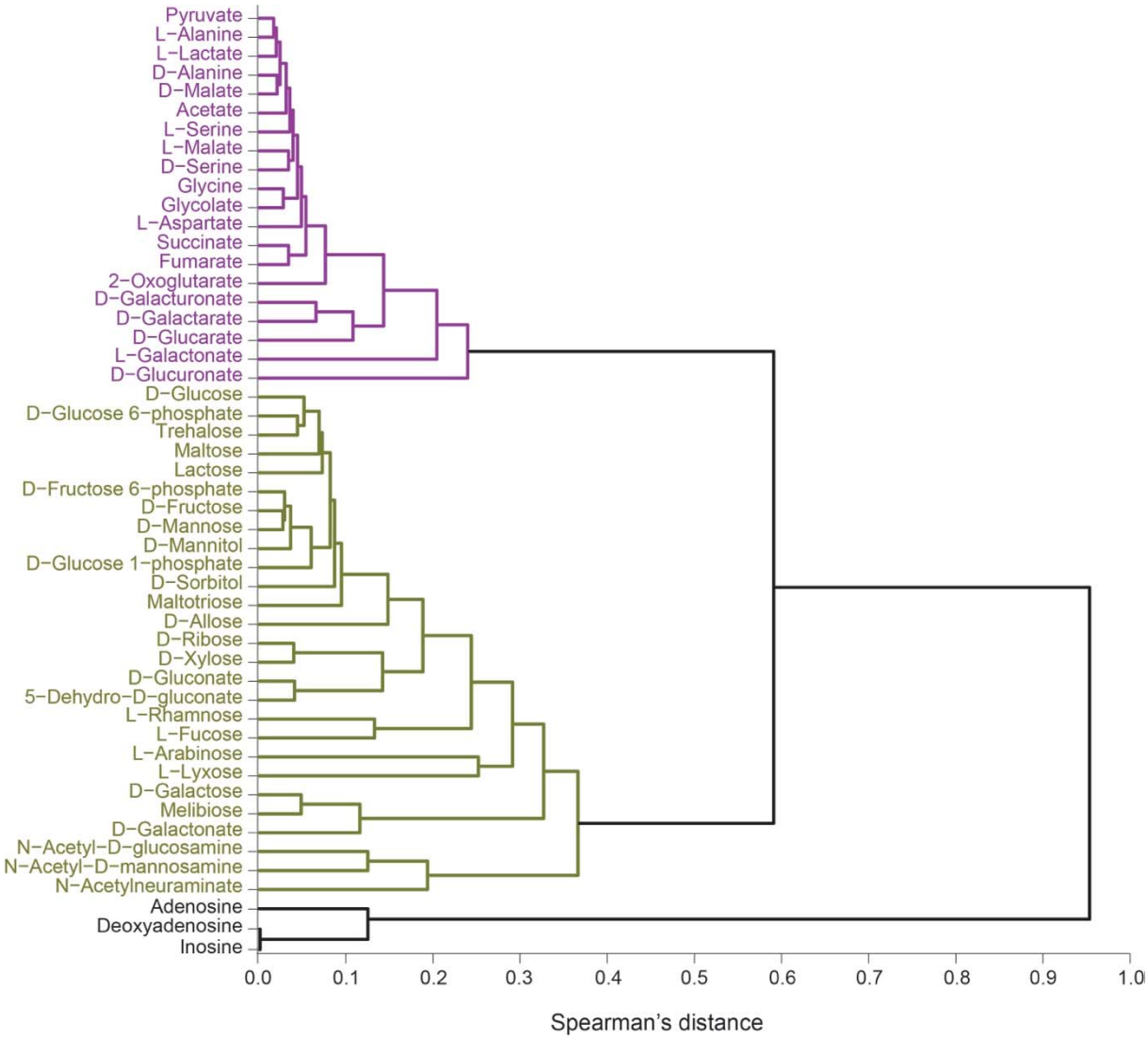
Viability required on C	$C_{j=1}$	$C_{j=2}$	$C_{j=3}$	$C_{j=4}$	$C_{j=5}$
$C_{i=1}$	1	0.3	0.05	0.8	0.57
$C_{i=2}$	0.5	1	0.06	0.14	0
$C_{i=3}$	0.01	0.68	1	0.7	0.03
$C_{i=4}$	0.99	0	0.2	1	0.32
$C_{i=5}$	0.1	0.23	0.43	0.09	1

b

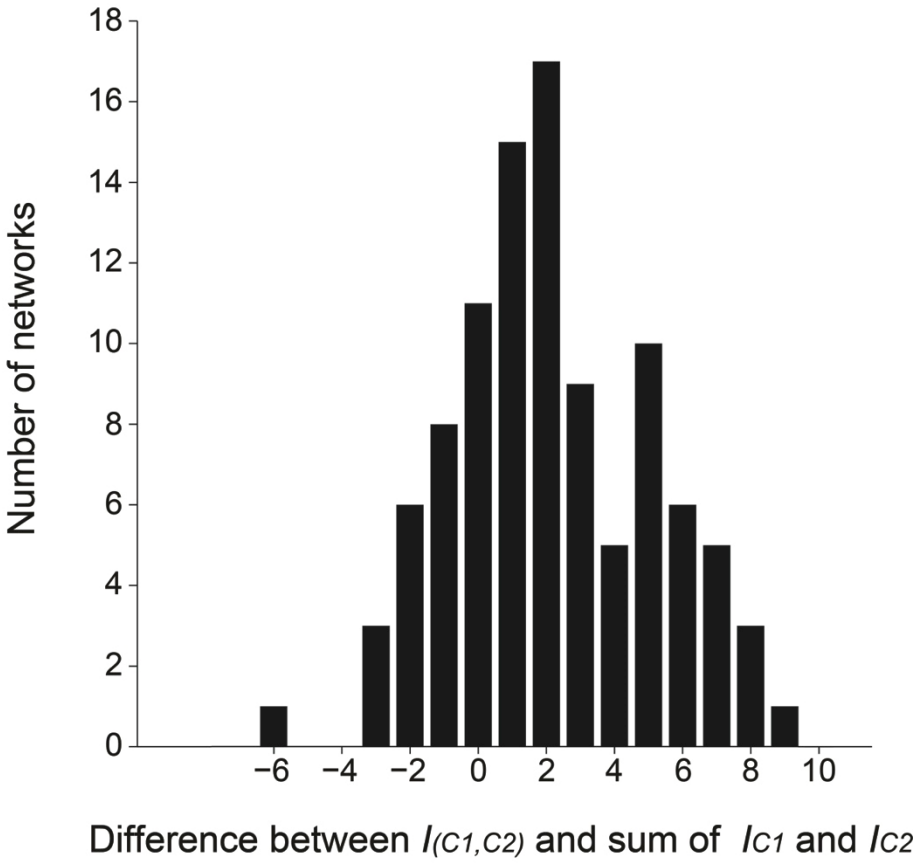


Supplementary figure S7: Innovation occurs preferentially within clusters of related carbon sources. (a) A hypothetical innovation matrix comprising 5 carbon sources. Each row vector corresponds to the carbon source C_i on which viability is required, and each column vector correspond to the additional

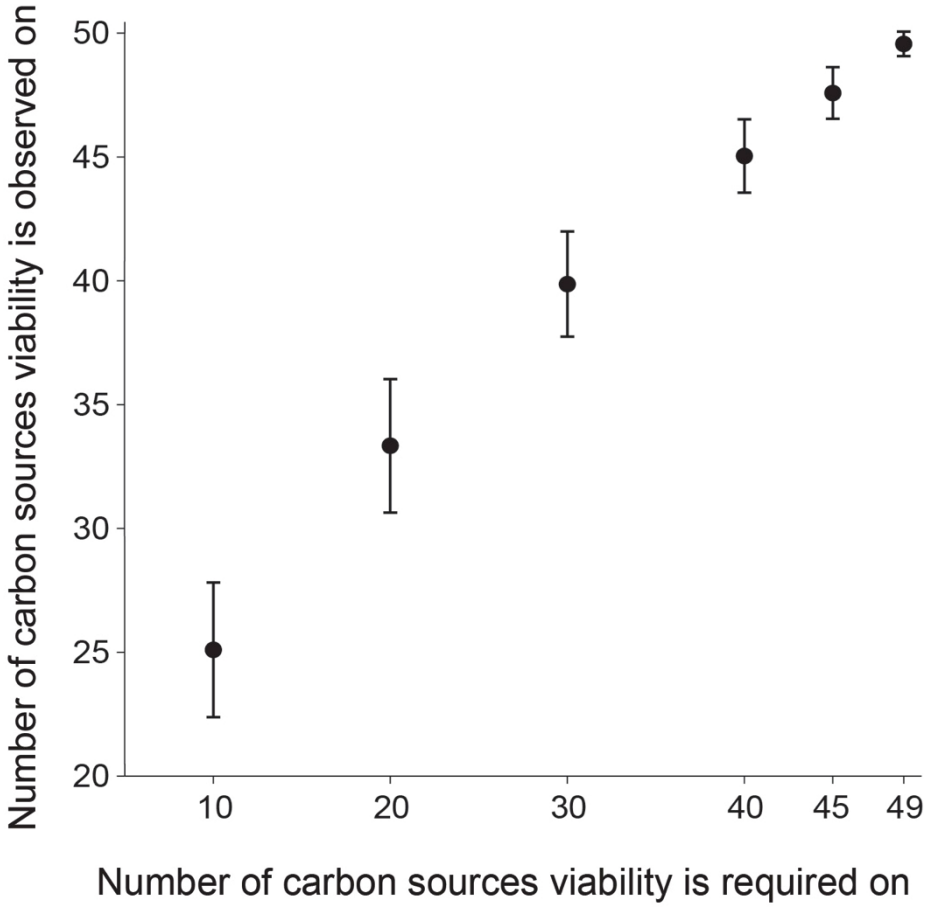
carbon source C_j. Each matrix entry indicates the fraction of networks that are also viable on C_j while required to be viable on C_i. (b) The figure shows a heat-map of the innovation matrix, organized according to different groups of carbon sources. The purple metabolite lettering corresponds to gluconeogenic carbon sources, green lettering corresponds to glycolytic carbon sources, and black corresponds to nucleotides as carbon sources. The two extreme ends of the colour spectrum of the heat map are blue and red, where blue (red) indicates that none (all) random networks required to be viable on carbon source C_i (rows) are also viable on an additional carbon source C_j (columns). The figure shows that carbon sources within a cluster favour the utilization of other carbon sources within the same cluster. Data in figures (a)-(b) are based on 50 samples of 500 random viable networks, where networks in each sample were required to be viable on a different source of 50 different carbon sources.



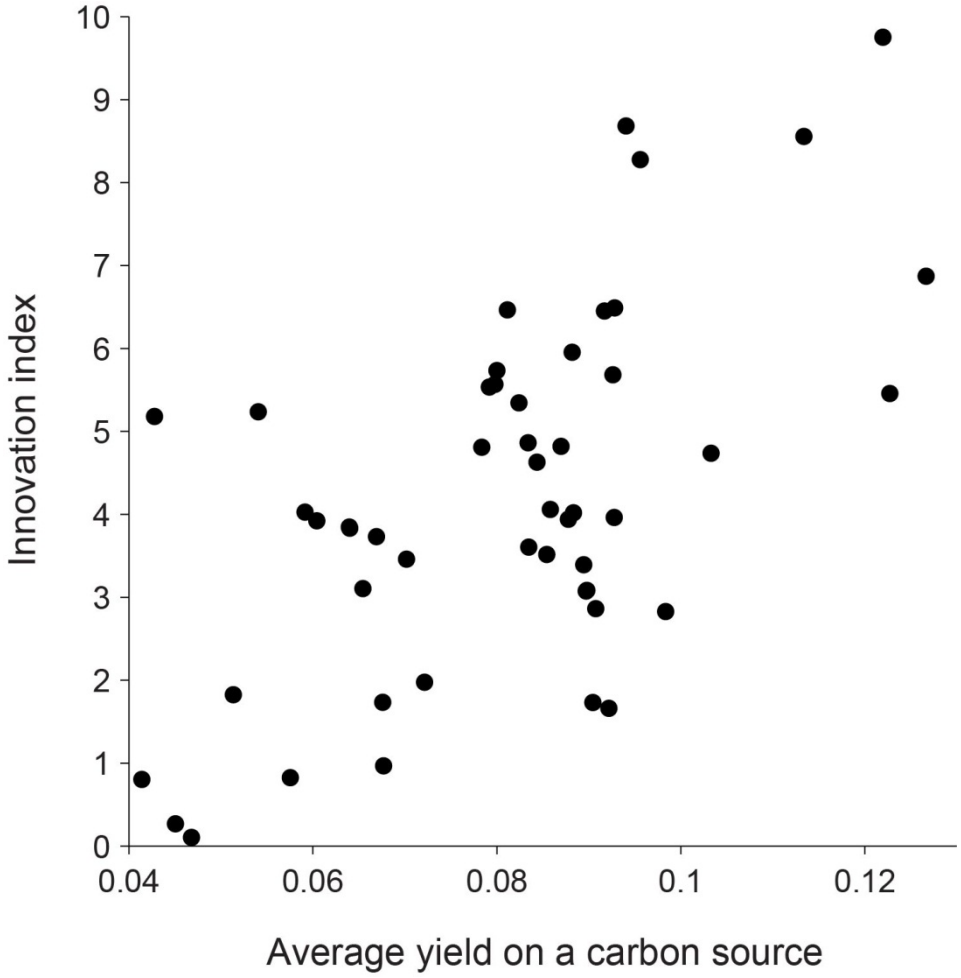
Supplementary figure S8: Innovation occurs preferentially within clusters of related carbon sources. The dendrogram shows three distinct groups of carbon sources based on hierarchical clustering of the innovation matrix using the Spearman's rank correlation distance (horizontal axis, see supplementary methods). The red, green, and blue groups of metabolites correspond to glycolytic, gluconeogenic, and nucleotide carbon sources. Note that the Spearman's distance between any two clusters of carbon sources is larger than 0.6. Data are based on 50 samples of 500 random viable networks, where networks in each sample were required to be viable on a different one of 50 different carbon sources.



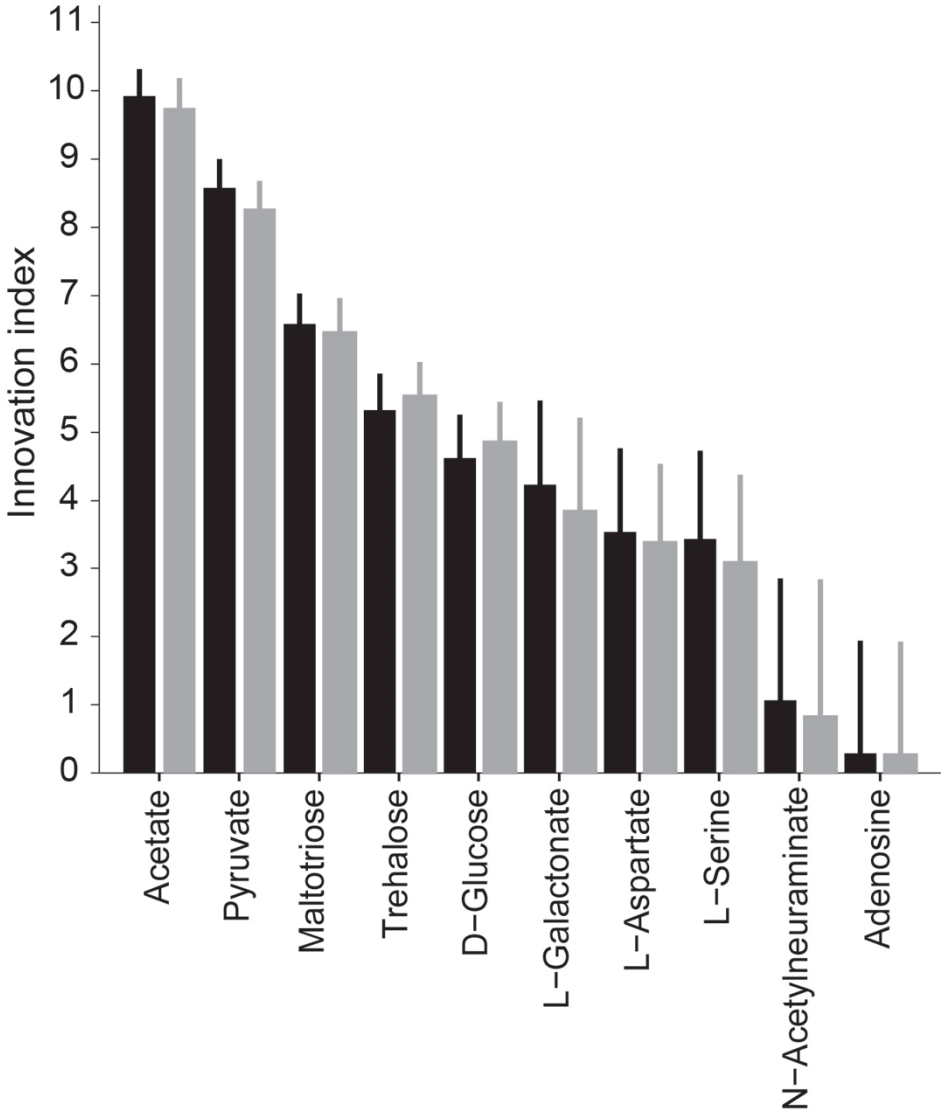
Supplementary figure S9: Pre-adaptation through required viability on two carbon sources is synergistic. The figure shows the distribution of the quantity $I_{(C1,C2)} - I_{C1} - I_{C2}$, averaged over 100 random metabolic networks viable on a pair of carbon sources C_1 and C_2 (horizontal axis). This quantity describes whether the innovation index of a pair of carbon sources ($I_{(C1,C2)}$) is higher or lower than the sum of the individual innovation indices I_{C1} and I_{C2}. A majority of pairs (77 percent) have a synergistic mean innovation index ($I_{(C1,C2)} > (I_{C1} + I_{C2})$), while the remaining pairs have an antagonistic innovation index ($I_{(C1,C2)} < (I_{C1} + I_{C2})$). Data are based on innovation vectors of 100 random networks viable on a pair of carbon sources (C_1, C_2), computed for 100 randomly chosen pairs of 50 carbon sources.



Supplementary figure S10: Diminishing returns in pre-adaptation. The vertical axis indicates the mean number of carbon sources on which viability is observed, for networks required to be viable on the number of randomly chosen carbon sources shown on the horizontal axis. For each value on the horizontal axis, data is based on specific samples of carbon sources, and on samples of 100 networks for each sample of carbon sources. Error bars denote one standard deviation. Note that networks are required to be viable on 49 carbon sources to allow viability on all 50 carbon sources studied here.



Supplementary figure S11: Innovation potential rises with reduced waste production. Each data point corresponds to one of 50 carbon sources. The horizontal axis indicates the average biomass yield per mole of carbon for the carbon source. The vertical axis indicates the average innovation index of the carbon source. Carbon sources that are efficiently metabolized (and produce low carbon waste) have a high yield. The figure shows that such high-yield carbon sources also allow viability on a greater number of additional carbon sources. For each carbon source, data are based on samples of 500 random networks viable on the carbon source (n = 25000).



Supplementary figure S12: A sample size of 500 networks is sufficient for our analysis. For each of 10 carbon sources *C* (horizontal axis), the figure indicates the mean innovation index (bar) and its coefficient of variation (lines) for 5000 random networks (black bars) and 500 random networks (gray bars) required to be viable on carbon source *C*. Note the broad distribution of the index. The height of the solid lines indicates the coefficient of variation. Note that the pairs of black and gray bars have similar height.

3.7 References

1. Darwin, C. On the Origin of Species. Charles Darwin. With an introduction by Ernst Mayr. Harvard University Press, Cambridge, Mass., 1964 (facsimile of the first edition, 1859). x 502 pp. 5.95. *Science (80-.).* **146,** 51–52 (1859).

2. Gould, S. J. & Vrba, E. S. Exaptation-a missing term in the science of form. *Paleobiology* **8,** 4–15 (1982).

3. True, J. R. & Carroll, S. B. Gene co-option in physiological and morphological evolution. *Annu. Rev. Cell Dev. Biol.* **18,** 53–80 (2002).

4. Zákány, J. & Duboule, D. Hox genes in digit development and evolution. *Cell Tissue Res.* **296,** 19–25 (1999).

5. Keys, D. N. *et al.* Recruitment of a hedgehog regulatory circuit in butterfly eyespot evolution. *Science* **283,** 532–4 (1999).

6. Tomarev, S. I. & Piatigorsky, J. Lens Crystallins of Invertebrates. Diversity and Recruitment from Detoxification Enzymes and Novel Proteins. *Eur. J. Biochem.* **235,** 449–465 (1996).

7. Pievani, T. & Serrelli, E. Exaptation in human evolution: how to test adaptive vs exaptive evolutionary hypotheses. *J. Anthropol. Sci.* **89,** 9–23 (2011).

8. Schuster, P., Fontana, W., Stadler, P. F. & Hofacker, I. L. From sequences to shapes and back: a case study in RNA secondary structures. *Proc. Biol. Sci.* **255,** 279–84 (1994).

9. Lipman, D. J. & Wilbur, W. J. Modelling neutral and selective evolution of protein folding. *Proc. Biol. Sci.* **245,** 7–11 (1991).

10. Cowperthwaite, M. C., Economo, E. P., Harcombe, W. R., Miller, E. L. & Meyers, L. A. The ascent of the abundant: how mutational networks constrain evolution. *PLoS Comput. Biol.* **4,** e1000110 (2008).

11. Ferrada, E. & Wagner, A. A Comparison of Genotype-Phenotype Maps for RNA and Proteins. *Biophys. J.* **102,** 1916–1925 (2012).

12. Samal, A., Matias Rodrigues, J. F., Jost, J., Martin, O. C. & Wagner, A. Genotype networks in metabolic reaction spaces. *BMC Syst. Biol.* **4,** 30 (2010).

13. Poulsen, T. S., Chang, Y.-Y. & Hove-Jensen, B. D-Allose Catabolism of Escherichia coli: Involvement of alsI and Regulation of als Regulon Expression by Allose and Ribose. *J. Bacteriol.* **181,** 7126–7130 (1999).

14. Meijnen, J.-P., de Winde, J. H. & Ruijssenaars, H. J. Engineering Pseudomonas putida S12 for efficient utilization of D-xylose and L-arabinose. *Appl. Environ. Microbiol.* **74,** 5031–7 (2008).

15. Feist, A. M. *et al.* A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol. Syst. Biol.* **3,** 121 (2007).

16. Neidhardt, F. & Ingraham, J. *Escherichia coli and Salmonella typhimurium: Cellular and Molecular Biology*. **1,** (American Society for Microbiology, Washington, DC, 1987).

17. Vieira-Silva, S. & Rocha, E. P. C. The systemic imprint of growth and its uses in ecological (meta)genomics. *PLoS Genet.* **6,** e1000808 (2010).

18. Nam, H. *et al.* Network context and selection in the evolution to enzyme specificity. *Science* **337,** 1101–4 (2012).

19. Aguilar, C. *et al.* Genetic changes during a laboratory adaptive evolution process that allowed fast growth in glucose to an Escherichia coli strain lacking the major glucose transport system. *BMC Genomics* **13,** 385 (2012).

20. Price, N. D., Reed, J. L. & Palsson, B. Ø. Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nat. Rev. Microbiol.* **2,** 886–97 (2004).

21. Nam, H. *et al.* Network context and selection in the evolution to enzyme specificity. *Science* **337,** 1101–4 (2012).

22. Kim, J., Kershner, J. P., Novikov, Y., Shoemaker, R. K. & Copley, S. D. Three serendipitous pathways in E. coli can bypass a block in pyridoxal-5'-phosphate synthesis. *Mol. Syst. Biol.* **6,** 436 (2010).

23. Martin, O. C. & Wagner, A. Multifunctionality and robustness trade-offs in model genetic circuits. *Biophys. J.* **94,** 2927–37 (2008).

24. Ancel, L. W. & Fontana, W. Plasticity, evolvability, and modularity in RNA. *J. Exp. Zool.* **288,** 242–83 (2000).

25. Amitai, G., Gupta, R. D. & Tawfik, D. S. Latent evolutionary potentials under the neutral mutational drift of an enzyme. *HFSP J.* **1,** 67–78 (2007).

26. Isalan, M. *et al.* Evolvability and hierarchy in rewired bacterial gene networks. *Nature* **452,** 840–5 (2008).

27. Kauffman, K. J., Prakash, P. & Edwards, J. S. Advances in flux balance analysis. *Curr. Opin. Biotechnol.* **14,** 491–6 (2003).

28. Goto, S., Nishioka, T. & Kanehisa, M. LIGAND: chemical database of enzyme reactions. *Nucleic Acids Res.* **28,** 380–2 (2000).

29. Goto, S., Okuno, Y., Hattori, M., Nishioka, T. & Kanehisa, M. LIGAND: database of chemical compounds and reactions in biological pathways. *Nucleic Acids Res.* **30,** 402–4 (2002).

30. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28,** 27–30 (2000).

31. Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M. & Hirakawa, M. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res.* **38,** D355–60 (2010).

32. Barve, A., Rodrigues, J. F. M. & Wagner, A. Superessential reactions in metabolic networks. *Proc. Natl. Acad. Sci. U. S. A.* **109,** E1121–30 (2012).

33. Matias Rodrigues, J. F. & Wagner, A. Evolutionary plasticity and innovations in complex metabolic reaction networks. *PLoS Comput. Biol.* **5,** e1000613 (2009).

34. Henry, C. S. *et al.* High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nat. Biotechnol.* **28,** 977–82 (2010).

35. Matias Rodrigues, J. F. & Wagner, A. Genotype networks, innovation, and robustness in sulfur metabolism. *BMC Syst Biol* **5,** 39 (2011).

36. Binder, K. & Heerman, D. W. *Monte Carlo Simulation in Statistical Physics*. (Springer, 2010).

37. Koskiniemi, S., Sun, S., Berg, O. G. & Andersson, D. I. Selection-driven gene loss in bacteria. *PLoS Genet.* **8,** e1002787 (2012).

38. Ochman, H., Elwyn, S. & Moran, N. A. Calibrating bacterial evolution. *Proc. Natl. Acad. Sci.* **96,** 12638–12643 (1999).

39. Wagner, A. & Fell, D. A. The small world inside large metabolic networks. *Proc. Biol. Sci.* **268,** 1803–10 (2001).

40. Ma, H.-W. & Zeng, A.-P. The connectivity structure, giant strong component and centrality of metabolic networks. *Bioinformatics* **19,** 1423–30 (2003).

41. Moore, E. The shortest path through a maze. in *Proc. Int. Symp. Theory Switch. Ann. Comput. Lab. Harvard Univ.* 285–292 (Harvard University Press, 1959).

42. Sokal, R. R. & Michener, C. D. A statistical method for evaluating systematic relationships. *Univ. Kansas Sci. Bull.* **28,** 1409–1438 (1958).

43. Pál, C., Papp, B. & Lercher, M. J. Horizontal gene transfer depends on gene content of the host. *Bioinformatics* **21 Suppl 2,** ii222–3 (2005).

44. Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. Equation of State Calculations by Fast Computing Machines. *J. Chem. Phys.* **21,** 1087 (1953).

45. Wovcha, M. G., Steuerwald, D. L. & Brooks, K. E. Amplification of D-xylose and D-glucose isomerase activities in Escherichia coli by gene cloning. *Appl. Envir. Microbiol.* **45,** 1402–1404 (1983).

46. Elias, M. D. *et al.* Occurrence of a bound ubiquinone and its function in Escherichia coli membrane-bound quinoprotein glucose dehydrogenase. *J. Biol. Chem.* **279,** 3078–83 (2004).

47. Vitkup, D., Kharchenko, P. & Wagner, A. Influence of metabolic network structure and function on enzyme evolution. *Genome Biol.* **7,** R39 (2006).

48. Hamming, R. W. Error detecting and error correcting codes. *Bell Syst. Tech. J.* **29,** 147–160 (1950).

4. Historical contingency does not strongly constrain the gradual evolution of metabolic properties in central carbon and genome-scale metabolisms

Aditya Barve[1,2], Sayed-Rzgar Hosseini[1,2,3], Olivier C Martin[4] and Andreas Wagner [1,2,5§]

[1] Institute of Evolutionary Biology and Environmental Sciences, Bldg. Y27, University of Zurich, Winterthurerstrasse 190, CH-8057, Zurich, Switzerland.

[2] The Swiss Institute of Bioinformatics, Bioinformatics, Quartier Sorge, Batiment Genopode, 1015, Lausanne, Switzerland.

[3] Computational Biology and Bioinformatics Master's Program, Department of Computer Science, ETH Zurich, Universitätsstrasse. 6, CH-8092, Zurich, Switzerland.

[4] INRA, UMR 0320/UMR 8120 Génétique Végétale, Univ Paris-Sud, F-91190 Gif-sur-Yvette, France.

[5] The Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA.

Part of this chapter was published in *BMC Systems Biology.* 8:48 (2014) [doi: 10.1186/1752-0509-8-48]

4.1 Abstract

A metabolism can evolve through changes in its biochemical reactions that are caused by processes such as horizontal gene transfer and gene deletion. While such changes need to preserve an organism's viability in its environment, they can modify other important properties, such as a metabolism's maximal biomass synthesis rate and its robustness to genetic and environmental change. Whether such properties can be modulated in evolution depends on whether all or most viable metabolisms – those that can synthesize all essential biomass precursors – are connected in a space of all possible metabolisms. Connectedness means that any two viable metabolisms can be converted into one another through a sequence of single reaction changes that leave viability intact. If the set of viable metabolisms is disconnected and highly fragmented, then historical contingency becomes important and restricts the alteration of metabolic properties, as well as the number of novel metabolic phenotypes accessible in evolution. We here computationally explore two vast spaces of possible metabolisms to ask whether viable metabolisms are connected. We find that for all but the simplest metabolisms, most viable metabolisms can be transformed into one another by single viability-preserving reaction changes. Where this is not the case, alternative essential metabolic pathways consisting of multiple reactions are responsible, but such pathways are not common. Metabolism is thus highly evolvable, in the sense that its properties could be fine-tuned by successively altering individual reactions. Historical contingency does not strongly restrict the origin of novel metabolic phenotypes.

4.2 Introduction

For biological systems on different levels of organization, the same broadly defined phenotype can usually be formed by more than one genotype. Examples include RNA, where many genotypes (sequences) share the same secondary structure phenotype [1–4]; proteins, where multiple amino acid sequences form the same fold [5,6]; regulatory circuits, where many genetically encoded circuit topologies can form the same expression pattern [7–9]; and metabolism, where multiple metabolic genotypes, encoding different combinations of chemical reactions, can confer viability on the same spectrum of nutrients [10–13]. The number of genotypes with the same phenotype is usually astronomical. For example, it can exceed 10^{20} for moderately long RNA molecules of 40 nucleotides with the same secondary structure [14]; it has been estimated at 10^{57} for proteins that adopt a fold characteristic of the bacteriophage λ transcriptional repressor [15], and at more than 10^{40} for model regulatory circuits of 10 genes that form a given gene expression pattern [7].

The many different genotypes that share one aspect of their phenotype may differ in other aspects, such as the thermodynamic stability of a given RNA or protein fold, the resilience of a gene expression pattern to stochastic noise, or the robustness of a metabolism to deletion of genes that encode metabolic enzymes [1,7,16,17]. Because such properties can be important for the biological function of any one system, the question whether they can be "fine-tuned" in evolution is important [7,18–20]. Such fine-tuning may depend on whether one can start from any one genotype with a given phenotypic property and reach most other such genotypes through sequences of small genetic change.

Whether such fine-tuning is possible can be studied in the framework of a space of possible genotypes, where two genotypes are adjacent if they differ by the smallest possible genetic change, such as a single amino acid change in two proteins. In this framework, the question becomes whether a set of genotypes with the same phenotype forms a single connected genotype network (also known as a neutral network [1]), or whether this network fragments into multiple isolated subnetworks or *disconnected components* [21].

Whenever such fragmentation occurs, the constraint it imposes on genotypic change does not only affect the ability to modulate a phenotype. It also gives an important role to historical accidents in the evolutionary process: The genotype with a given phenotype that evolution happened to have "discovered" first can determine the number and identity of other genotypes reachable through gradual genetic change. And by restricting the number of accessible genotypes, fragmentation can also restrict the spectrum of novel phenotypes accessible as new adaptations. The reason is that this spectrum depends strongly on a genotype's location in genotype space [22]. The further evolution can "walk away" from a given genotype, the more the spectrum of accessible phenotypes changes [1,11,23–26]. In sum, fragmentation of a genotype network can cause historical contingency and restrict a system's potential for future evolutionary change.

Existing work, based on computational models of phenotype formation, shows that fragmentation is system-dependent. For example, in RNA secondary structure phenotypes, genotype networks are typically highly fragmented [18,27], whereas for regulatory circuits, such fragmentation depends on the kind of circuit studied, its size, and how one defines its gene expression phenotypes [7,8,28]. Because the question has thus far not been answered in metabolic systems, we here analyze the connectedness of a space of metabolisms.

A metabolism is a complex network of chemical reactions, catalyzed by enzymes and encoded by genes, whose most fundamental task is to synthesize multiple small molecule precursors for biomass, such as amino acids, nucleotides, and lipids [29,30]. An organism's metabolic genotype is the part of a genome that encodes metabolic genes. It is thus fundamentally a string of DNA, but can be represented more compactly as a binary vector of length N, where N is the number of metabolic reactions in a known "universe" of metabolic reactions (supplementary figure S1 [10,11]). This universe comprises all enzyme-catalyzed reactions known to take place in some organism. The i-th entry of this vector corresponds to the i-th reaction in a list of such reactions, and for any one organism, the value of this entry is one if the organism can catalyze the i-th reaction, and zero otherwise. On evolutionary time scales, the reaction complement

of a metabolism can change through processes such as horizontal transfer of enzyme-coding genes, gene deletions, as well as gene duplications followed by sequence divergence.

The known "universe" of metabolism currently comprises more than $N = 5000$ reactions [31,32]. This means that there are more than 2^{5000} different metabolic genotypes, which constitute a vast space of possible metabolisms. For any one metabolism in this space and any one chemical environment, one can compute the spectrum of biomass precursors that it can synthesize using the constraint-based computational method of flux-balance analysis (FBA). We call any one metabolism *viable* in a specific chemical environment, if it can synthesize every single one in a spectrum of essential biomass precursors from nutrients in this environment [10,11,13,33] (see Methods). We will here consider minimal chemical environments that contain only one carbon source, such as glucose, as the sole carbon source.

Because connectedness of a metabolic genotype network may depend on the number n of reactions in a metabolism, we distinguish in our analysis metabolisms of different sizes. If $\Omega(n)$ is the set of all metabolisms with n biochemical reactions ($n \leq N$) and if $V(n)$ is the subset of all viable metabolisms, we are interested in whether $V(n)$ is connected. Because the metabolisms of free-living heterotrophic metabolisms may have thousands of reactions, we need to study $V(n)$ for metabolisms this large. This is not an easy task, because the set of viable metabolisms is so enormous that exhaustive enumeration is impossible [10,34]. Therefore, to sharpen our intuition and to illustrate key concepts, we first analyze a smaller metabolic genotype space whose viable metabolisms can be enumerated exhaustively. This is the space of metabolisms that can be formed by subsets of $N = 51$ reactions in central carbon metabolism [35] (see methods, section 4.5). Even though central carbon metabolism is highly conserved, its reaction complement varies in nature, for example through variants of glycolysis [36–40] and the tricarboxylic acid cycle, where some organisms have an incomplete cycle [41]. We go beyond such naturally occurring variation and analyze metabolisms comprised of all possible subsets of all 51 reactions. Even though this number of metabolisms is astronomical ($2^{51} \approx 10^{15}$), we were able to determine viability for all of them, and thus analyze the connectivity of $V(n)$ for all $n \leq N$ ($N = 51$). After that, we turn to larger,

genome-scale metabolisms, where we study the connectivity of $V(n)$ through a sampling approach.

Our observations show that for all but the simplest metabolisms, those that contain close to the minimal number of reactions necessary for viability, most viable metabolisms $V(n)$ lie on a single connected genotype network. Where fragmentation into different components occurs, its biochemical cause are alternative biochemical pathways that occur in different components, that are essential for the synthesis of specific biomass precursors, that comprise more than one reaction, and that cannot be transformed into one another by changes in single reactions without destroying viability. Because such pathways only occur in the smallest metabolisms, fragmentation and thus historical contingency do not strongly constrain the evolution of properties such as robustness, biomass synthesis rate, or the accessibility of novel metabolic phenotypes.

4.3 Results

Study System 1: Central Carbon Metabolism

Our first analysis focuses on metabolic genotypes that can be formed with subsets of $N = 51$ reactions in the central carbon metabolism of *E. coli* [35] (see methods, section 4.5). This metabolic core of *E. coli* includes reactions from glycolysis/gluconeogenesis, the tricarboxylic acid cycle, oxidative phosphorylation, pyruvate metabolism, the pentose phosphate shunt, as well as some reactions from glutamate metabolism (supplementary table S1a, section 4.6). It produces 13 precursor molecules (supplementary table S1b, section 4.6) that are required to synthesize all 63 small biomass molecules of *E. coli*, including nucleotides, amino acids, and lipids [30,35,42]. Examples of these precursors include oxaloacetate, a metabolite participating in the tricarboxylic acid cycle, which is used in the synthesis of amino acids such as asparagine, aspartate, lysine, and threonine [29,42]. Another example is ribose-5-phosphate, which participates in the pentose phosphate pathway, and is necessary for the synthesis of nucleotides and amino acids, such as histidine,

phenylalanine, and tryptophan [29,42]. In our analysis, we consider a metabolism *viable* only if it can synthesize all 13 of these biomass precursors in a well-defined minimal environment containing a specific *sole* carbon source, such as glucose (see methods, section 4.5).

The fraction of viable genotypes is extremely small and decreases as metabolism size *n* decreases

For each $n \leq N = 51$, we here explore the space $\Omega(n)$ of metabolisms (metabolic genotypes) with a given number of n reactions. We represent each such metabolism as a binary vector of length $N = 51$, whose i-th entry is equal to one if the i-th reaction is present and zero otherwise. The largest metabolism ($n = N$) is the one where all reactions are present. The space of all possible metabolisms that contain a subset of these 51 reactions has 2^{51} ($\approx 10^{15}$) member genotypes, while for a given n, $\Omega(n)$ contains $\binom{51}{n}$ genotypes. We are especially interested in the subset $V(n)$ of $\Omega(n)$ that consists only of viable metabolisms. Because, $\Omega(n)$ can be very large, determining $V(n)$ is no small undertaking. For example, for metabolisms with $n = 30$, $\Omega(n)$ contains more than 1.14×10^{14} genotypes, and the viability of each of them cannot be determined by brute force. However, one can use some peculiarities of metabolism to render this computation feasible (see methods, section 4.5). For example, consider a metabolism (the "parent") with n reactions and another metabolism (the "child") derived from it by deleting one reaction. If the parent is not viable then the child will not be viable either. By analyzing the viability of metabolisms with decreasing numbers of reactions n, and taking advantage of this relationship, we were able to reduce the computational cost of enumerating viable metabolisms by a factor $\approx 10^6$ to the evaluation of viability for only 1.55×10^9 metabolisms [43].

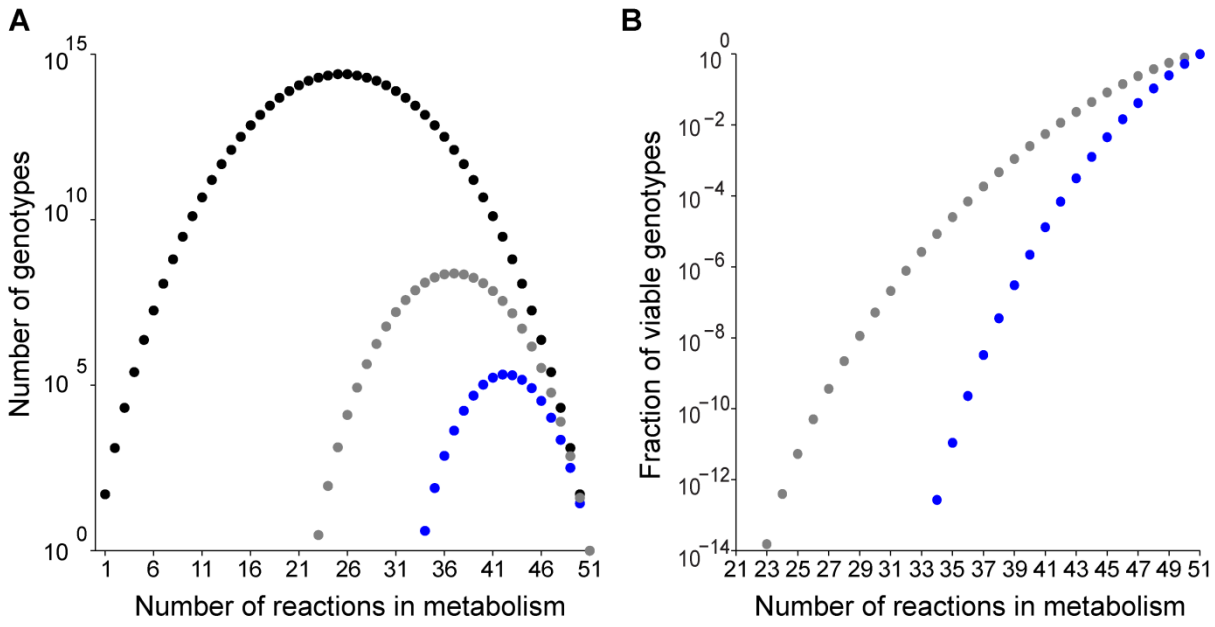


Figure 4.1: (a) The vertical axis (note the logarithmic scale) shows the number of genotypes, and the horizontal axis shows the number n of reactions in a metabolism. Black circles represent the number of genotypes in genotype space $\Omega(n)$ (regardless of viability), grey circles show the number of metabolisms viable on glucose, whereas the blue circles denote the number of metabolisms viable on all 10 carbon sources. (b) The vertical axis (note the logarithmic scale) shows the fraction $|V(n)|$ /$|\Omega(n)|$. The grey circles show the fraction of genotypes viable on glucose relative to the number of possible metabolisms, whereas the blue circles denote the fraction of genotypes viable on 10 carbon sources relative to the number of possible metabolisms. Note that viable genotypes become extremely rare as the number of reactions in a metabolism decreases. Data for both figures is based on all viable metabolisms for each n (supplementary table S2 and supplementary table S3, section 4.6).

Figure 4.1A shows the number of viable metabolisms $V(n)$ (grey circles), together with the number of all metabolisms (black circles, $\Omega(n) = \binom{51}{n}$) as a function of the number n of reactions. Note the logarithmic vertical axis. The absolute number of viable metabolisms has a maximum at $n = 37$ with a total of 2.39×10^8 metabolisms, while the minimum size of a viable metabolism, i.e., the smallest n such that $V(n) > 0$ is 23 (supplementary table S2, section 4.6). This means that at least 23 reactions are required to synthesize all 13 biomass precursors on glucose. There are three such smallest metabolisms, one of which is shown in supplementary figure S2 (section 4.6). Figure 4.1B expresses $V(n)$ as a fraction of the number of metabolisms $\Omega(n)$ (grey circles), and shows that this fraction decreases with decreasing n. This means that random sampling is much less likely to yield a viable metabolism for small than

for large metabolisms. For the smallest n with viable metabolisms (n = 23), the three viable metabolisms correspond to a fraction 10^{-14} of all metabolisms of size 23. The largest viable metabolism contains all n = 51 reactions.

Useful principles to determine the connectedness of genotype networks.

The viable genotypes at any one size n can be represented as a genotype network, a graph whose nodes are genotypes, and where two genotypes are adjacent (connected by an edge), if they share all but one reaction. For example, the two hypothetical genotypes G_1 and G_2, where G_1 consists of reactions $\{R_1, R_2, R_3\}$, and G_2 consists of reactions $\{R_2, R_3, R_4\}$, are adjacent. This is because G_1 and G_2 share two out of the three reactions (R_2 and R_3). One can reach G_2 from G_1 by adding reaction R_4 and removing R_1, an event that we refer to as a reaction swap [10,12,33]. This definition of neighboring genotypes allows us to keep the number of reactions in a genotype network constant. We note that each reaction swap can be decomposed into the addition of a reaction followed by the deletion of a reaction, both of which preserve viability provided that the reaction swap does. In other words, genotype networks that are connected if adjacency is defined under reaction swaps will remain connected if adjacency is defined via a sequence of alternating reaction additions and reaction deletions.

Our principal goal is to identify whether genotype networks at any one size n are connected. This first requires us to establish the adjacency of $\binom{V(n)}{2}$ genotype pairs, followed by application of standard graph theory algorithms such as breadth-first search [21,44] to compute whether genotypes decompose into two or more disconnected components, or whether they form a single connected network, i.e., whether a path through $V(n)$ exists connecting any two genotypes [21]. Because $V(n)$ exceeds 10^6 genotypes at intermediate n (supplementary table S2, section 4.6), such conventional methods lead to large computational cost for all but the largest and smallest metabolisms (n = 23-28 and n = 46-50 reactions). For genotype networks comprising metabolisms of intermediate size (n = 29-45), we therefore took advantage of another relationship between “parent” and “child” metabolisms, namely that the connectivity

of a genotype network at size n can be understood based on its connectivity at size n-1. We explain this relationship next.

Starting from a genotype $G(n)$ with n reactions, one can obtain a parent genotype $G(n+1)$ with $(n+1)$ reactions by adding to it any one reaction among the $N = 51$ reactions that are not already part of $G(n)$. Because addition of a reaction does not eliminate viability, $G(n+1)$ will be viable, and thus be a member of $V(n+1)$. For any one genotype $G(n)$, there exist N-n reactions that are not part of this genotype. Therefore, one can obtain exactly N-n genotypes of size $n+1$ by adding a single reaction to a genotype $G(n)$. And because each pair of these genotypes of size $n+1$ shares all but one reaction (the newly added reaction), every parent genotype in this set is adjacent to every other parent genotype. In other words, these genotypes form a clique in $V(n+1)$ [21].

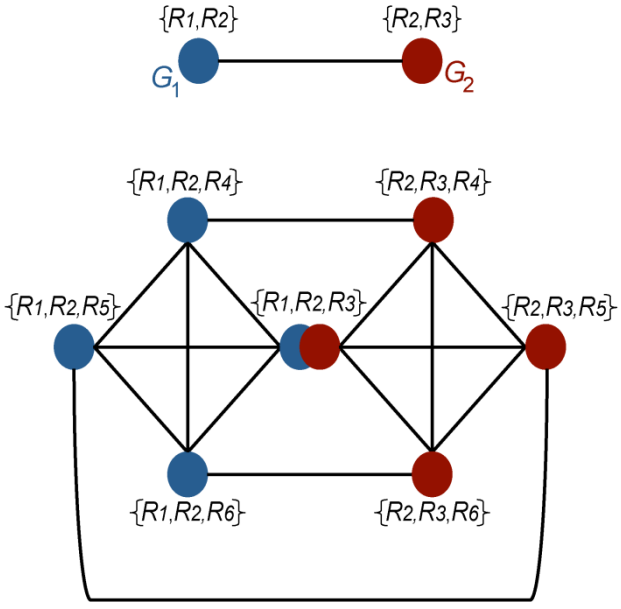


Figure 4.2 - Connectivity of metabolisms can be inferred from parent and child relationship. The figure uses a hypothetical example of two neighboring metabolisms with three reactions each (upper panel) to illustrate the relationship between the connectedness of genotypes with n reactions ($G(n)$) and their "parents" of $n+1$ reactions that can be obtained from them by adding a single reaction (lower panel). Importantly, if genotypes $G(n)$ form a connected set, then all genotypes $G(n+1)$ obtained by adding one reaction to each of them also form a connected set.

We next point out that if two genotypes of size n are adjacent, then their corresponding genotypes of size $(n+1)$ form two cliques linked by at least one genotype of size $(n+1)$. The hypothetical example in figure 4.2 illustrates this fact. Consider a "universe" of only $N = 6$ reactions - $\{R_1, R_2, R_3, R_4, R_5, R_6\}$. The upper half of figure 4.2 shows two hypothetical genotypes (G_1 in blue and G_2 in red) that are

viable, adjacent, and contain two reactions each ($n = 2$). Genotype G_1 comprises reactions $\{R_1, R_2\}$, while the other genotype G_2 comprises reactions $\{R_2, R_3\}$. The lower part of the figure shows all genotypes containing three reactions each that can be obtained from adding one reaction to genotypes G_1 and G_2. Blue genotypes are parents of G_1, whereas red genotypes are parents of G_2. Note that the red and blue genotypes form two cliques. Among the 7 genotypes of size n+1 that are parents of either G_1 or G_2, one is special, because the two cliques share it. In our example, this is the genotype containing reactions $\{R_1, R_2, R_3\}$. More generally, this shared genotype is the one genotype obtained from a pair of adjacent genotypes G_1 and G_2 in $V(n)$ by adding the reaction to G_1 that it does not share with G_2, or vice versa. (There is only one such reaction, because G_1 and G_2 are adjacent.) We note that additional edges connect genotypes in both cliques (figure 4.2). Specifically, those edges connect those genotypes derived from adding the same reaction to G_1 and G_2. There are exactly (N-n-1) such edges.

These observations have the following important corollary: If a genotype network containing genotypes of size n is connected, then all genotypes $G(n+1)$ obtained from genotypes of size n are also connected.

Number of reactions in a metabolism (*n*)	**Number of components**	**Number of viable metabolisms *V*(*n*)**	**Number of minimal metabolisms**	**Fraction of minimal metabolisms**
23	2	3	3	1
24	2	91	8	0.08791
25	2	1333	23	0.01725
26	2	12512	14	0.00111
27	1	84344	27	0.00032

28	2	434238	43	9.9×10^{-5}
29	1	1773969	28	1.57×10^{-5}
30	1	5900578	15	2.54×10^{-5}

Table 4.1 - The number of minimal metabolisms in genotype networks from the central carbon metabolism. The left-most column shows the number of reactions *n* in a metabolism, the second column from the left shows the number of disconnected components into which the genotype network of these viable metabolisms fragments, the third column shows the number of viable metabolisms for each *n*, and the fourth column shows the number of minimal metabolisms. Note that the fraction of minimal metabolisms (column five) decreases as metabolism size *n* increases.

So far, our line of reasoning explains connectedness of genotypes that are parents of connected genotypes at a lower size. But some viable genotypes are not parents of any other genotype. These are exactly those genotypes in which elimination of any one reaction abolishes viability. We have called such genotypes *minimal* [10,12], and note that they do not necessarily correspond to the smallest metabolisms. For example, there are 8 metabolisms that are viable on glucose and that have 24 reactions, all of which are essential (table 4.1), but the smallest viable metabolisms on glucose have only 23 reactions. (We explain further below why minimal metabolisms may vary in their number of reactions.). As one increases the size of a metabolism, such "childless" metabolisms could in principle arise at any *n*. Since our preceding argument about connectedness does not apply to them, they need to be identified, and their connectedness to the rest of a genotype network needs to be examined separately, as discussed in the next section.

We identified minimal metabolisms at each size *n* by deleting every single reaction from each genotype in $V(n)$, and by examining whether the resulting genotype was viable, and thus identifying those genotypes in which no reaction can be deleted. Table 4.1 shows the number of minimal metabolisms at each *n*, and demonstrates that their proportion among all viable metabolisms ($V(n)$) decreases dramatically with increasing metabolism size *n*. Importantly for the next section, no minimal metabolisms viable on glucose exist above $n = 30$.

In sum, we here observed that if a genotype network is connected at size *n*, the genotype network formed by the parents of its genotypes is also connected. Because minimal metabolisms are not parents of any other metabolisms, they need to be analyzed separately.

Metabolic genotype networks are connected for all but the smallest metabolisms

To determine connectedness of genotype networks for metabolisms *V(n)* viable on glucose, we began by analyzing the smallest (*n* = 23-28) and largest (*n* = 46-50) metabolisms. We did so by computing, first, edge lists for each genotype network, and, second, the connectedness of the genotype network, using the graph analysis software `igraph` [45]. We found that viable metabolisms of size *n* = 27, as well as *n* = 46 to *n* = 50 have only one connected component. In contrast, viable metabolisms of sizes *n* = 23, 24, 25, 26, and 28 are fragmented. They form a genotype network with two components (table 4.2).

Number of reactions in a metabolism (*n*)	**Number of viable metabolisms *V(n)***	**Number of components**	**Fraction of viable metabolisms in the large connected component**
23	3	2	0.667
24	91	2	0.637
25	1333	2	0.997
26	12512	2	0.992
27	84344	1	1
28	434238	2	0.977

Table 4.2 - Fragmentation occurs in central carbon metabolisms close to minimal number of reactions. For each metabolism size, the table shows the number of viable metabolisms, the number of components, and the fraction of genotypes in the largest component. The genotype network comprising metabolisms of size 23 contains three metabolisms, of which two form one component and the other is isolated from them. For metabolisms of size 24, the two components are almost of the same size, and the larger component contains 63.74 percent of the viable genotypes. For larger metabolisms, the

genotype network is largely connected, with more than 97 percent of genotypes belonging to the largest component.

Fragmented genotype networks may decompose into components with different sizes, such that the majority of genotypes belong to the largest component. In this case, most viable genotypes can be reached from each other through a series of small genotypic changes that affect only single reactions each and that leave the phenotype constant. Alternatively, fragmentation of a genotype network may result in components with similar size, which can impede accessibility of many genotypes. Table 4.2 shows that this is not generally the case. The fourth column denotes the fraction of genotypes belonging to the largest component of a genotype network, and it shows that the largest components of the genotype networks at size n = 25-28 encompass almost all (i.e., more than 99 percent) of the viable genotypes. At n = 23, the genotype network has two components that consist of one and two metabolisms. At size n = 24 there are 91 viable genotypes, 58 (64 percent) of which belong to the largest component.

We next turn to a more detailed analysis of genotype network fragmentation at the smallest sizes. Figure 4.3 below shows graph representations of genotype networks whose metabolisms have sizes n = 23, 24 and 25. Filled circles represent genotypes. Adjacent genotypes are connected by an edge. The size of a circle corresponds to the number of neighbors of the corresponding genotype. Minimal metabolisms are shown in red in all three panels. All three metabolisms of size 23 are minimal (figure 4.3A). Two of them are adjacent metabolisms and form component *A* (left), whereas the remaining isolated metabolism forms component *B* (right). The green and orange circles of figure 4.3B show the result of adding one reaction from the remaining pool of 28 reactions (N-n = 51 - 23 = 28) to each of the two genotypes in component *A*. Such addition yields a connected component *A'* of 55 metabolisms with 24 reactions (green, figure 4.3B). The component consists of two cliques connected to each other by a single connected genotype. Analogous addition of reactions to the single genotype in component *B* of figure 4.3A yields a connected component *B'* with 28 metabolisms (orange) of size 24. The total number of genotypes in component *A'* and *B'* is 83. However, there are 91 viable metabolisms with size 24 (Table 4.1). It turns out that the missing eight metabolisms are minimal (red) and cannot be derived using

reaction addition to metabolisms at size 23. Three of them are connected to component *A'* and five of them to component *B'* (figure 4.3B). Overall, the number of components at size 24 reflects the number of components at size 23, because these components are derived from the smaller components at size 23. This, however, is no longer true for the genotype network of metabolisms with 25 reactions in figure 4.3C. In this panel, genotypes shown in green (components *A''*) and orange (components *B''*) are parents of the green and orange genotypes in components *A'* and *B'* respectively. Notice that these components are now connected, in contrast to their disconnectedness at size 24. What connects them are some of the minimal metabolisms that arose anew at size 24 and 25. There are 23 such child-less minimal genotypes at size 25 (figure 4.3C and table 4.1). Four of them form a new component labeled *C* (center bottom of figure 4.3C).

An analogous analysis of metabolisms up to size 30 can help understand why all larger metabolisms must be connected (supplementary text S1, supplementary figure S3, section 4.6). There are two germane observations. First, at size $n = 30$, there are approximately 5.9×10^6 metabolisms and all of them fall into a single connected component (supplementary figure S1 (section 4.6), and table 4.1). Second, no minimal metabolisms exist at size 31 and beyond (table 4.1). This means that all parent metabolisms at size $(n+1)$ are derived from child metabolisms at sizes beyond $n = 30$. By our argument in the preceding section, they must therefore form a single connected component (figure 4.2).

In sum, we showed that genotype networks formed by different central carbon metabolism variants are connected in metabolic genotype space for all but the smallest viable metabolisms. With few exceptions, wherever fragmentation occurs, more than 99 percent of genotypes belong to the largest component. This high connectivity arises from the parent-child relationships we discussed (figure 4.2), as well as from the relatively small number of minimal metabolisms that arise at each n (table 4.1).

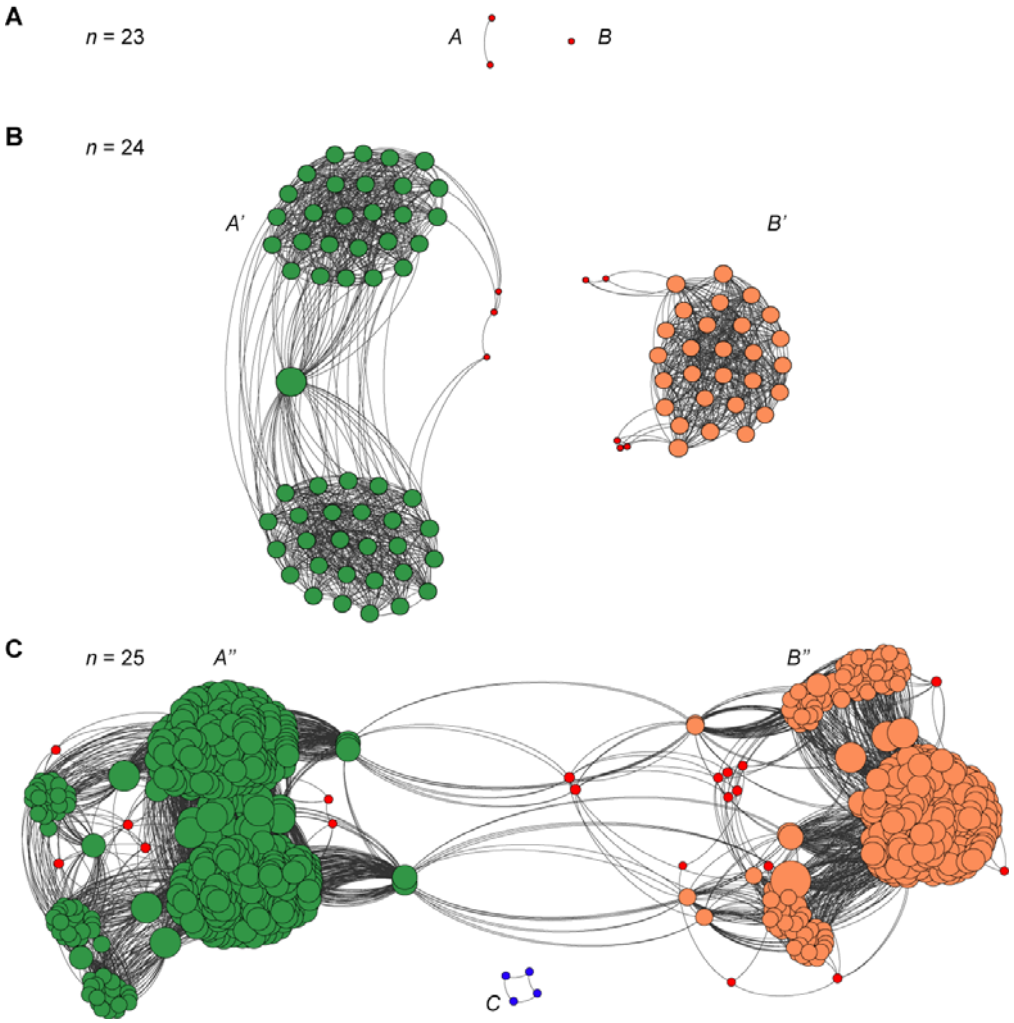


Figure 4.3 - The figure shows the genotype networks of metabolisms containing 23 (A), 24 (B) and 25 (C) reactions. Each filled circle corresponds to a genotype. Two genotypes are connected by an edge (curved line) if they are neighbors. Red circles correspond to minimal metabolisms of a given number of reactions *n*. The size of each circle corresponds to its number of neighboring genotypes, which increases as metabolism size increases. Graphs were drawn using the graph visualization software Gephi [46].

Essential pathways cause genotype network fragmentation

Thus far, our analysis focused on broad patterns of genotype network fragmentation. We next discuss the possible mechanistic reasons for such fragmentation. They revolve around different biochemical pathways that are essential for viability among metabolisms in different components. Essential reactions are those whose removal

results in a loss of viability (see methods, section 4.5), and a reaction's essentiality may depend on other reactions present in a metabolism. That is, a reaction can be essential in one metabolism, but nonessential in another metabolism, because of the presence of alternative metabolic routes [13]. The fraction of metabolisms of a given size in which a reaction is essential is a useful quantifier of the reaction's essentiality, which we have called the reaction's superessentiality index [13]. The concept of (super)essentiality can be extended to entire metabolic pathways, groups of essential reactions that share substrates/products with each other and cannot be replaced without a loss of viability.

We next illustrate with an example how pathway (super)essentiality causes fragmentation of genotype networks, by demonstrating the existence of alternative essential pathways in different network components for metabolisms with 23 and 24 reactions. To identify such pathways, we first computed the superessentiality index of reactions in metabolisms of size 23 and 24 each, and did so for all genotypes in each of the two genotype network components (figure 4.3A and 4.3B) separately (see methods, section 4.5). We then examined which reactions differ in their superessentiality index between the two components. We found five such reactions, which can be subdivided into groups of two and three reactions, respectively. The first group comprises the reactions catalyzed by transketolase 1 (TKT1) and transaldolase (TALA). They are essential in all metabolisms from network component *A'* (figure 4.3), but inessential in all metabolisms belonging to component *B'*. The second group comprises the reactions catalyzed by the enzymes glucose-6-phosphate dehydrogenase (G6PDH), 6-phosphogluconolactonase (PGL), and phosphogluconate dehydrogenase (GND). They are essential in all metabolisms of component *B'*, but inessential in any of the genotypes in component *A'* (figure 4.3). Taken together, this means that TKT1 and TALA form a small but essential pathway in the genotypes belonging to component *A'*, while G6PDH, PGL and GND form another essential pathway in genotypes belonging to component *B'*.

These five reactions are part of the pentose phosphate pathway, as shown in figure 4.4 below. The pentose phosphate pathway is required for the synthesis of two biomass precursors, ribose-5-phosphate (r5p) and erythrose-4-phosphate (e4p) (solid squares

in figure 4.4). The reactions shown in black are essential in metabolisms belonging to both genotype network components (figure 4.3A and 4.3B). In contrast, the essentiality of reactions participating in the two alternative essential pathways (green and orange), which contain the reactions discussed in the preceding paragraph, depends on which of the two components a metabolism belongs to. To understand why, we first note that the metabolites glucose-6-phosphate (g6p), fructose-6-phosphate (f6p), and glyceraldehyde-3-phosphate (g3p) are also synthesized by reactions in glycolysis, and thus constitute metabolic inputs to the pentose phosphate pathway for the synthesis of e4p and r5p.

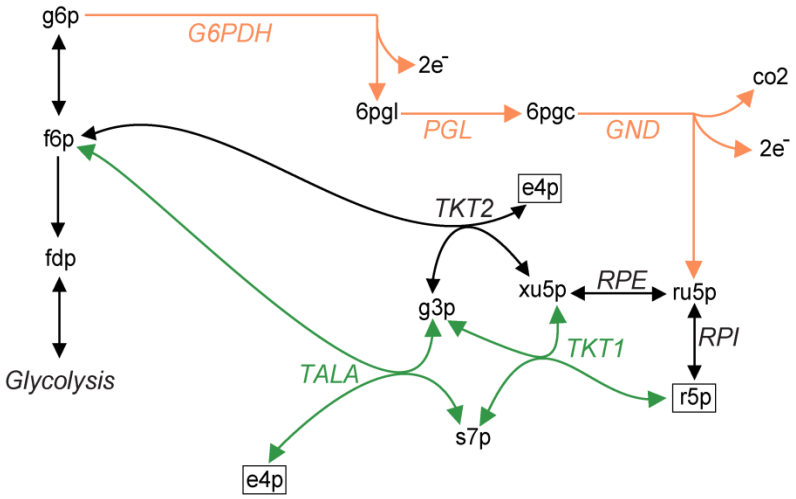


Figure 4.4 - Essential pathways in pentose phosphate metabolism. The figure shows the two essential pathways (orange and green) in pentose phosphate metabolism that are necessary for the synthesis of biomass precursors e4p and r5p (in square boxes). Reactions in black are essential regardless of the metabolism in which they occur. Reactions catalyzed by G6PDH, PGL and GND form an essential pathway (orange), while reactions catalyzed by TALA and TKT1 (green) form another essential pathway. If the reactions in orange are absent, the reactions in green become essential and vice versa. This is because removal of TALA and TKT1 requires the synthesis of r5p through the reactions catalyzed by G6PDH, PGL and GND, while removal of these reactions forces the synthesis of r5p through TAL and TKT1. Note that metabolites g6p, f6p and g3p also participate in glycolysis and therefore can be produced there and supplied to the pentose phosphate pathway. Enzymes catalyzing each of the reactions are shown in uppercase italic typeface. Abbreviations - g6p, D-glucose-6-phosphate; r5p, D-ribose-5-phosphate; e4p, D-erythrose-4-phosphate; f6p, D-fructose-6-phosphate; fdp, fructose-diphosphate; g3p, glyceraldehyde-3-phosphate; G6PDH, glucose-6-phosphate dehydrogenase; PGL, 6-phosphogluconolactonase; GND, 6-phosphogluconate dehydrogenase; RPI, ribose-5-phosphate isomerase; RPE, ribose-5-phosphate 3-epimerase; TKT1, transketolase 1; TALA, transaldolase; TKT2, transketolase 2.

Flux balance analysis can be used to show that reactions catalyzed by transketolase 1 (TKT1) and transaldolase (TALA) are required to synthesize sufficient r5p for viability (supplementary table S1 - biomass reaction) upon removal of any one reaction from the orange pathway (G6PDH, PGL, GND), thus rendering the reactions catalyzed by TKT1 and TALA essential. Conversely, removal of any one reaction from the green pathway (TKT1, TALA) leads to a requirement for all reactions in the orange pathway to produce the pathway output. In sum, the genotypes of size 23 and 24 are disconnected because alternative essential pathways exist in them that consist of more than one essential reaction, and because no one reaction in one pathway can replace a reaction in the other pathway. Put differently, loss of any one reaction in one pathway can only be compensated by addition of all reactions of the other pathway. Because metabolisms at size 23 are separated by three swaps, genotype space can be connected at size 25 (subgraphs *A''* and *B''*), that is, after successive addition of two reactions.

In the supplementary results ('Essential pathways cause genotype network fragmentation' in section 4.6) we discuss another example, which illustrates that essential and alternative metabolic routes need not contribute to biosynthesis of the same precursors, and may arise in functionally different and unrelated parts of metabolism. These differences notwithstanding, the examples illustrate the mechanistic reason for genotype network fragmentation: It is not possible to interconvert two genotypes in different components by one reaction swap because such interconversion will inevitably create inviable genotypes in which two alternative essential pathways are incomplete.

As a corollary, the longer such essential alternative pathways are, the greater the number of reactions m that need to be added to non-adjacent viable genotypes $G(n)$, such that viable genotypes $G(n+m)$ become connected.

Metabolisms viable on multiple carbon sources are also mostly connected

Many organisms are viable on multiple carbon sources, which may impose additional constraints on a metabolism. We wished to find out how severely these constraints affect genotype network connectivity in our analysis of central carbon metabolism. To

this end, we analyzed metabolisms that are a subset of our $N = 51$ reactions and that are viable on a total of 10 common carbon sources when each of them is provided as the sole carbon source (see methods for all these carbon sources in section 4.5). Because glucose is among these 10 carbon sources, metabolisms viable on all 10 carbon sources are also viable on glucose. In other words, the genotype network they form at any specific metabolism size n is a subset of the genotype network of metabolisms viable on glucose. We note that the metabolism comprising all $N = 51$ reactions is viable on all 10 different carbon sources.

We used an approach identical to that described above for glucose to identify metabolisms viable on all 10 carbon sources. Their numbers are shown in figure 4.1A (blue circles), which shows that, first, no metabolism with fewer than $n = 34$ reactions is viable on all 10 carbon sources, whereas the minimal size is much smaller (n=23) for metabolisms viable on glucose alone (table 4.1, supplementary tables S2 and S3, section 4.6). Second, the number of metabolisms viable on 10 carbon sources is much smaller than the number of metabolisms viable on glucose. It reaches a maximum at $n = 42$ with 2.1×10^5 metabolisms (supplementary table S3), many fewer than for viability on glucose (2.39×10^8 metabolisms at $n = 37$). This difference is also highlighted in figure 4.1B whose vertical axis represents genotypes viable on glucose (grey) and on all ten different carbon sources (blue) as a fraction of all genotypes. At the minimal size of $n = 34$ reactions, genotypes viable on all 10 different carbon sources comprise approximately one $10^{-8\text{th}}$ of those genotypes viable on glucose of the same size.

This strong constraint on metabolisms viable on multiple carbon sources raises the possibility of genotype network fragmentation. However, we found no evidence for such fragmentation. Because the number of genotypes viable on 10 carbon sources is relatively small, we were able to use standard algorithms to determine their connectedness, which show that genotype networks of all sizes except for $n = 35$ and 36 reactions consist of only one connected component. At size $n = 35$ the genotype network fragments into three components. The largest of them contains 91.13 percent of viable genotypes. At size $n = 36$, the network fragments into two components, with the larger containing 99.6 percent of genotypes. This implies that one can access any

metabolic genotype viable on 10 carbon sources, regardless of its size, from most other viable genotypes through a series of individual reaction changes.

Study system 2: Genome-scale metabolisms

We have thus far studied connectedness for metabolisms drawn from the reduced reaction set of central carbon metabolism, which comprises a small subset of the more than 1000 reactions in the typical metabolism of a free-living organism. In this section, we focus on the connectedness of larger, genome-scale metabolisms. Their reactions come from the known "universe" of possible biochemical reactions, which comprises, at our present state of partial knowledge, already more than 5000 reactions [31,32]. For any one such metabolism to be viable, we require that it is able to synthesize all 63 essential biomass precursors of *E. coli* [30] – most of which are molecules central to all life, such as nucleotides and amino acids (see methods, section 4.5) – in a minimal environment containing glucose as the sole carbon source.

Using our binary representation of a metabolic genotype, the number of possible genome-scale metabolisms is greater than 2^{5000}, which renders exhaustive analysis of connectivity infeasible. Random sampling of the space using Markov Chain Monte Carlo (MCMC) methods can be very useful [10,11,13,33], but it is not suitable for our purpose, because the MCMC approach samples genotypes from the same component of a genotype network.

We thus use a different sampling approach [10,12,33,47], which starts from a "global" metabolism that comprises all reactions in the known universe (and is viable on glucose). This metabolism has 5906 reactions. Its viable children would form a single connected component, but as one reduces their number of reactions further, the set of viable genotypes *V(n)* might become disconnected. Figure 4.5A illustrates this possibility schematically. It shows a funnel-like landscape whose width at a given number of reactions *n* (vertical axis) indicates the number of viable metabolisms at this *n*. The number of viable metabolisms approaches zero as *n* approaches the smallest possible size at which a metabolism can be viable. Starting from the global metabolism, one can randomly select a sequence of reactions for deletion while

requiring that each deletion retain viability. Parts of three hypothetical deletion sequences are shown as three trajectories in the panel. Two of them (solid) lead into deep depressions in the funnel, which correspond to disconnected components of a genotype network. More precisely, a metabolism that resides in one such depression cannot be converted into another viable metabolism without changing its number of reactions (the altitude in the landscape), as doing so would require it to traverse the exterior of the funnel. The third trajectory (dotted line) enters such a depression only at a much lower number of reactions. We wanted to know whether such funnels appear in the landscape at moderate *n* (figure 4.5A) or only at values of *n* close to the smallest number of reactions permitting viability (figure 4.5B).

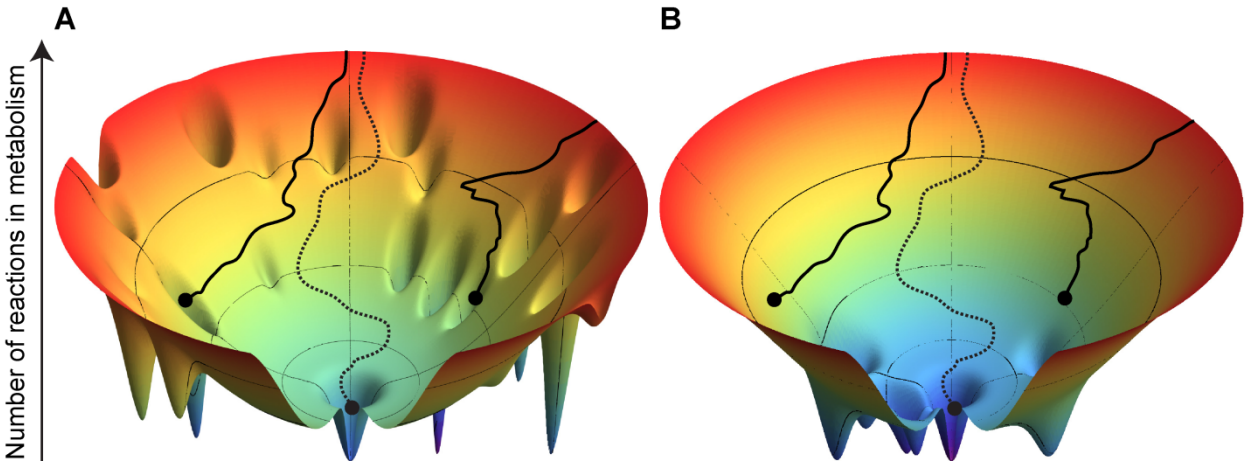


Figure 4.5 - A spatial schematic of genotype network connectivity at different metabolism sizes *n*. Each panel shows a funnel-like landscape, where the funnel's width reflects the number of metabolisms that are viable at any given number *n* of reactions (altitude in the landscape). The lowest points of the landscape correspond to metabolisms whose size are close to the smallest possible *n* needed for viability. Solid and dotted lines denote random sequences of reaction removal from some viable starting metabolism (not shown) that terminate at a particular size *n* and result in some viable metabolism denoted by circles. Depressions in the funnel correspond to disconnected components of the genotype network. Disconnected components in (A) arise at higher metabolism sizes than in (B). Two of the trajectories in A (solid lines) terminate in such depressions, that is, in metabolisms that are part of a disconnected component. These metabolisms are not interconvertible through viability preserving reaction swaps. The same trajectories in B terminate at a part of the funnel where all viable metabolisms are still connected. The figure was generated using a script from (http://www.oaslab.com/Drawing_funnels.html)

To find out, we derived multiple viable metabolisms with a given size *n* as follows. Starting from the global metabolism, we repeatedly deleted randomly chosen reactions from it, such that each deletion preserved viability, until we had arrived at a minimum metabolism, that is, a metabolism whose number of reactions cannot be reduced further. In doing so, we kept track of the deleted reactions, and the sequence in which they were deleted. Each minimum metabolism created in this way had fewer than 400 reactions (see also below). We used these minimal metabolisms, as well as information about the sequence in which reactions were deleted, to create larger metabolisms of varying sizes *n*, each of which corresponds to a specific point in the deletion sequence. We repeated this procedure 500 times, which allowed us to create 500 minimal metabolisms, as well as 500 metabolisms of various intermediate sizes.

Most viable genome-scale metabolisms reside in the same connected component

If genotype networks were highly fragmented at a given size *n*, then different random deletion sequences would yield metabolisms that reside in different components of a genotype network. In this case, it would not be possible to connect two metabolisms that reside in different components of a genotype network through a sequence of reaction swaps, each of which preserves viability. With these observations in mind, we attempted to connect metabolisms in our samples of viable metabolisms of a given size (see Methods). Specifically, for any sample of metabolisms [G_1, G_2, G_3, … , G_{500}], we attempted to connect G_i and G_{i+1} ($1 \leq i < 500$) through viability-preserving reaction swaps. We did this for 500 metabolisms of size 1400 (similar to that of *E. coli*), 1000, 500, and 400 (above the size of minimal metabolisms, see below). In this way, we were able to show that all 500 metabolisms are connected at each of these sizes. Thus, down to a size of $n = 400$ reactions, the genotype network of metabolisms viable on glucose is not highly fragmented, and one component comprises the vast majority or all metabolisms.

Because many free-living microorganisms are viable on multiple carbon sources, we generated 500 additional metabolisms through the reaction deletion process just described, but with the additional constraint that they remain viable on ten sole carbon

sources (the same ten as used in our analysis of central carbon metabolism). Specifically, we created again metabolisms of size 1400, 1000, 500, 450 and 425 (slightly above the size of minimal metabolisms for viability on 10 carbon sources). We then repeated the procedure that attempts to connect genotypes G_i and G_{i+1} through viability-preserving reaction swaps. In this way, we were able to show that all 500 metabolisms are connected at each of these sizes. Thus, down to a size of $n = 425$ reactions, the genotype network of most metabolisms viable on all 10 carbon sources consists of one connected component.

It is possible to make this point more quantitatively and establish a statistical bound on the fraction of metabolisms contained in the largest connected component of $V(n)$. Specifically, let us consider the null hypothesis that more than one percent of $V(n)$ resides outside this largest component. If this null hypothesis is correct, then the probability p that a randomly drawn viable genotype is *not* on this largest component is greater than $p = 0.01$. Moreover, the probability that some number M of genotypes drawn at random from $V(n)$ all fall on the largest connected component would be smaller than $(1-p)^M$. In our case, $M = 500$ and $(1-p)^M < 0.99^{500} = 0.0066$. In other words, the results of our samplings allow us to reject the above null hypothesis at a significance level smaller than 1 percent.

Minimal metabolism size can help explain connectedness

In the sections on central carbon metabolism we showed that new components disconnected from the remainder of a genotype network can arise as one increases metabolism size, and that they originate from "childless" minimal metabolisms which appear at a given size n that is small compared to the total number of possible reactions. To examine their size for larger metabolic system, we studied the 500 minimal metabolisms that we derived from the sequential random deletion strategy described in the previous section (figure 4.6A). Their size ranges from 324 to 391 reactions, with a mean of 352 reactions (standard error = 11.44 reactions) (figure 4.6A). Although we cannot absolutely exclude the possibility that minimal metabolisms exist with more than 400 reactions, the fact that all of the minimal metabolisms we found have fewer reactions suggests that the emergence of new

genotype network components will be rare above 400 reactions. This observation further supports our assertion that most metabolisms with more than 400 reactions will be part of a single genotype network. It also means that essential alternative metabolic pathways of more than one reaction that are characteristic for a given connected component exist only for small metabolisms. Alternative pathways for the synthesis of most biomass molecules undoubtedly exist, but most of them can be converted into one another through sequences of single reaction changes that preserve viability.

In a final analysis, we asked whether the minimal networks that our approach identified are isolated in metabolic genotype space $\Omega(n)$, or whether they might themselves form large components. To this end, we simply asked whether these networks have any viable neighbors in $\Omega(n)$, metabolisms that differ by a single reaction swap, which are also viable. The result (figure 4.6B) shows that even minimal metabolisms have typically hundreds of neighbors. Specifically, an average minimal metabolism has 372.8 viable neighbors (standard deviation: 79 neighbors). The maximum number of neighbors for a minimal metabolism is 685. Figure 4.6C shows that larger minimal metabolisms tend to have more neighbors than smaller ones (Spearman's $\rho = 0.43$, p-value $< 10^{-22}$). Taken together, this means that minimal metabolisms themselves must form large components and are certainly not isolated. It mirrors the situation in central carbon metabolism, where newly emerging minimal metabolisms at a given size n also form connected components, albeit small ones (figure 4.3).

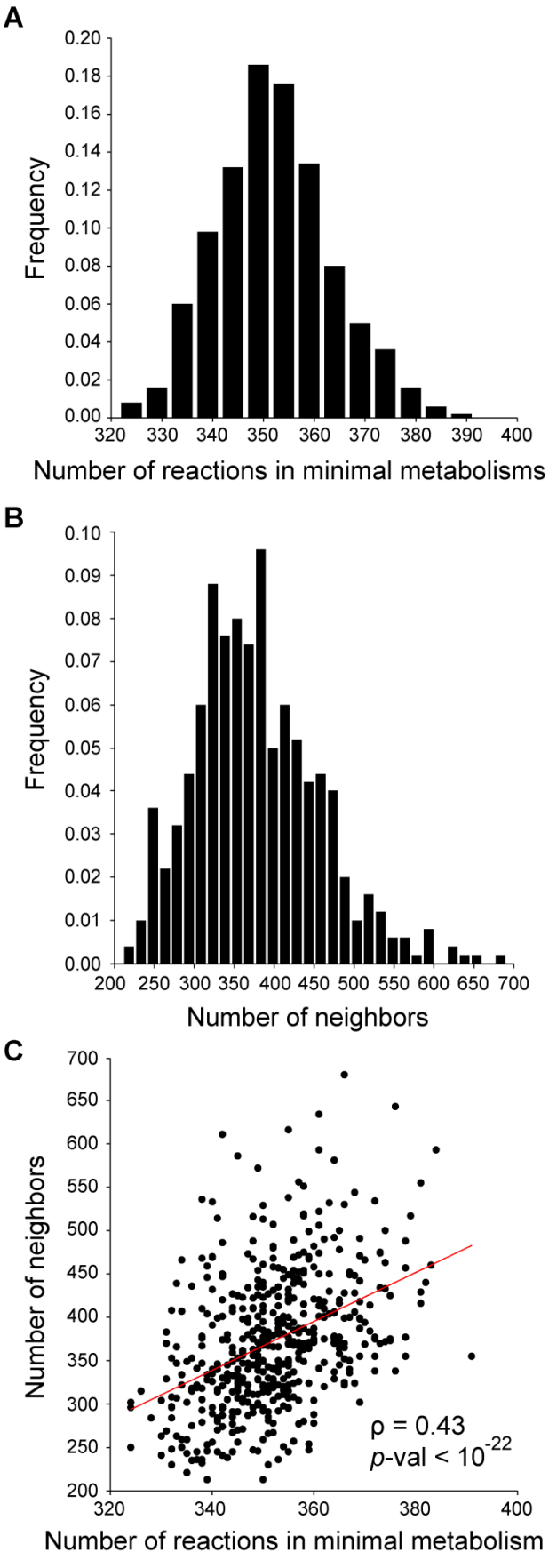


Figure 6 - Minimal metabolisms from the complete universe can have many viable neighbors. (A) The horizontal axis denotes the size of minimal metabolisms and the vertical axis denotes their frequency. The average minimal metabolism comprises of 352 reactions, while the largest minimal metabolism we find has 391 reactions. (B) The vertical axis shows the frequency of metabolisms with a given number of neighbors (horizontal axis). A minimal metabolism has 372.8 viable neighbors on average. Data in (C) show that the number of viable neighbors (vertical axis) is positively correlated with the number of reactions present in a minimal metabolism (horizontal axis). Data in (A), (B) and (C) are based on 500 minimal metabolisms generated through the random reaction deletion process described in the text.

4.4 Discussion

To our knowledge, our analysis of all possible ≈10^{15} metabolisms comprising subsets of reactions in central carbon metabolism is the first exhaustive analysis of a metabolic space this large, even though smaller-scale analyses were carried out before with different goals [48–50]. For instance, Ebenhöh *et. al*., studied metabolisms comprising only reactions that catalyzed change in the carbon skeleton. They found that metabolisms that can synthesize specific products (have the same phenotype) tend to be connected through single reaction changes and tend to cluster in close neighborhoods[51]. Our analysis focused on metabolisms viable on glucose, which are required to synthesize 13 products of central carbon metabolism that are biomass precursors. We found that viable metabolisms could have fewer than half (23) of the maximal number of 51 reactions in central carbon metabolism. Moreover, for metabolisms covering 77 percent of the size of the viable range (n = 23 - 51), all (n = 29 - 51) or the vast majority of metabolisms of size n form a single connected component (network) in the space of metabolisms.

In genome-scale metabolisms, where exhaustive enumeration is no longer possible, and where we required the synthesis of 63 common biomass molecules for viability, we found viable metabolisms with as few as 324 reactions, and for 93.23 percent of the size range of viable metabolisms (n = 400 - 5906) the vast majority of metabolisms form a single connected component of a genotype network. More specifically, with a probability of greater than 0.99, 99 percent of all viable metabolisms with more than 400 reactions are part of the same component. We note that it would have been sufficient to perform the sequential reaction deletion procedure needed to arrive at this conclusion for metabolisms of size n = 400, and not also for metabolisms of size n = 400-1400, as we did. The reason is an elementary observation we made about metabolic genotype space: If a set of viable metabolisms $V(n)$ is connected for some number of reactions n, then $V(m)$ must be connected for all $m > n$, provided that no new minimal metabolisms appear at any value of m. The largest minimal metabolism we found has n = 391 reactions, and while we cannot exclude the existence of minimal metabolisms above n = 400 with certainty, such metabolisms would be increasingly rare at large n. They would create new genotype

network components that would comprise a vanishing fraction of the rest of the connected genotype network (even though they might contain many metabolisms in absolute numbers).

Figure 4.5B illustrates schematically the dependence of fragmentation on metabolism size *n* that we observed. Depressions in the funnel-like landscape whose width reflects the number of viable metabolisms correspond to disconnected metabolic networks and appear only at small altitudes (metabolism sizes). That is, the hypothetical landscape of figure 4.5B reflects our observations, whereas that of figure 4.5A, where disconnected metabolisms appear at much higher reaction numbers does not.

With possible exceptions in some marine bacteria [52,53] metabolisms with sizes as small as $n = 400$ are not usually found in free-living organisms. They occur in (endo)symbionts [54,55] and (endo)parasites [56,57], which live in close association with a host organism and are provided nutrients and a constant environment which allows them to shed many enzyme-coding genes [58–61,47]. Organisms that have lived inside a host for a long time experience less of the kinds of evolutionary change – especially horizontal gene transfer – that is powerful in endowing the genomes of free-living organisms with new evolutionary adaptations [58,60]. In other words, the fragmentation of genotype networks that we see for very small metabolisms, and that can constrain their evolution, is of little relevance for the evolution of free-living organisms. Those organisms whose evolution it could constrain the most are already subject to little evolutionary change for ecological reasons.

Our analysis of genotype network fragmentation provides a coarse, statistical view on the organization of genotype space. This view needs to be complemented by a mechanistic perspective that asks what distinguishes the metabolisms that exist in different components of a genotype network. What could prevent evolution from converting them into each other through a series of single viability-preserving reaction changes? The answer lies in alternative metabolic pathways that are essential for the biosynthesis of one or more biomass molecules. Metabolisms in one genotype network component have one such pathway, and metabolisms in the other component have another such pathway. (In addition, these metabolisms may differ in other

essential pathways.) At least one of these pathways must comprise more than one reaction; otherwise the two metabolisms could be converted into one another through a single reaction swap. We have provided two examples, one involving the biosynthesis of erythrose-4-phosphate and ribose-5-phosphate through variants of the pentose phosphate pathway, the other concerning the biosynthesis of phosphoenolpyruvate.

For two reasons, such alternative essential pathways are not likely to hamper the evolution of most metabolic systems. First, we observed fragmentation only for relatively small metabolisms, which means that in larger metabolisms, alternative *essential* pathways with more than one reaction do not exist. They can usually be converted into each other by single reaction changes that do not cause a loss of viability. Second, our analysis required that we impose change through reaction swaps – a reaction addition paired with a deletion – that leave reaction numbers constant. However, this is not usually how evolutionary change in a metabolism's reactions occurs. For example, horizontal gene transfer frequently adds more than one gene and thus more than one reaction to a metabolism [62–64]. However, our approach of single reaction swaps also allows us to understand evolution of metabolism with respect to a single reaction change (unit change). Given that the genotype spaces are extremely vast, single reaction changes increase the probability of finding viable genotypes, thus allowing us to understand connectedness in the simplest sense. Also, in a metabolism that harbors one of two alternatives for an essential pathway, a horizontal gene transfer event may introduce the genes of the other pathway. After that, the two pathways may coexist, and the first pathway is free to deteriorate through loss of function mutations in its genes. A potential example of co-existing alternative pathways involves the two pathways responsible for synthesizing isopentenyl diphosphate (the 1-deoxy-D-xylulose 5-phosphate pathway and the mevalonate pathway), a molecule that is required for the synthesis of isoprenoids. Some actinomycetes that harbor both pathways in their complete forms may have obtained the responsible genes through horizontal gene transfer [65]. In sum, common forms of genetic change can help bridge different components of a genotype network where such components exist.

The main limitation of our work comes from the enormous computational cost associated with evaluating the viability of many metabolisms. While our sampling approach for genome-scale metabolisms allowed us to circumvent this problem for any one carbon source, it is possible that viability on a broader range of carbon sources (or sources of other chemical elements) might have led to greater genotype network fragmentation. This possibility is suggested by our analysis of central carbon metabolism, where metabolisms viable on 10 carbon sources must have at least 34 reactions, and fragmentation of genotype networks stops at 37 reactions. However, for genome-scale metabolisms, the metabolism sizes at which fragmentation would cease would increase only modestly with each additional carbon source on which viability is required. This is because previous work has shown that viability on every additional carbon source requires on average the addition of only two reactions to a metabolism [66]. For example, viability on ten additional carbon sources would increase the size of minimal metabolisms by only 20 reactions. Because the number of minimal metabolisms that arise *de novo* with increasing metabolic complexity n is closely linked to the metabolism size at which fragmentation occurs, viable genotype networks $V(n)$ would still remain connected over the vast majority of the range of n. Indeed, we found that genome-scale metabolisms viable on 10 carbon sources and comprising 425 reactions are connected in genotype space and belong to the same component. Possible exceptions might involve metabolisms viable on hundreds of different carbon sources, but even environmental generalists are typically not viable on that many. (The generalist *E. coli* is viable on some 50 alternative carbon sources [30].

Another limitation of our work is that we only considered viability on carbon sources. We cannot exclude the possibility that viability on sources of different chemical elements may lead to different fragmentation patterns. However, it is unlikely that carbon sources are exceptional in this regard. For example, the minimal size of metabolisms viable on different sulfur sources comprises only 90 reactions, and is thus even smaller than that of metabolisms viable on carbon [12]. The reason is that fewer biomass molecules contain sulfur, an observation that also holds for the two other key elements nitrogen and phosphorus.

A further limitation is that we focus on evolutionary constraints caused by the presence or absence of biochemical reactions, rather than on differences in the regulation of existing enzymes or their encoding genes. Such regulatory constraints can influence important metabolic properties such as biomass growth rate [20]. However, they can also be easily broken through regulatory evolution, even on the short time scales of laboratory evolution experiments [20,19,67]. Reaction absence is thus a more fundamental constraint, but we note that the exploration of regulatory constraints remains an important task for future work.

Finally, we do not consider one potential cause of genotype network fragmentation: If one required for viability that biomass precursors need to be synthesized at a high rate, then genotype networks may fragment more often than we observe. However, fast biomass synthesis and its main consequence, rapid cell division, are not universally important outside the laboratory environment. For example, a survey of microbial growth rates shows that many microbes have very long generation times in the wild [68]. Rapid growth thus may not be a biological sensible requirement for viability in many wild organisms.

In sum, our analysis has shown that over a broad range of metabolic complexity, historical contingency is not likely to constrain the modulation of metabolic properties, or the accessibility of novel metabolic phenotypes. Only the smallest metabolisms, which typically do not occur in free-living organisms, are likely to be subject to such constraints, which stem from genotype network fragmentation. Additional factors that we did not consider explicitly are likely to further reduce such fragmentation. For instance, promiscuous enzymes are highly ubiquitous in metabolism, which are capable of catalyzing more than one biochemical reaction [69,70]. Such reactions may only increase the number of functional pathways and potentially increase connectedness between disconnected components. Additionally, horizontal gene transfer events may add multi-reaction metabolic pathways to an existing metabolism, and thus bridge otherwise disconnected genotype network components.

4.5 Methods

Flux balance analysis.

Flux balance analysis (FBA) is a constraint-based computational method [34,71] that can predict synthetic abilities and other properties of metabolisms – complex networks of enzyme-catalyzed biochemical reactions. Any one such network can comprise anything from a few dozen reactions, such as central carbon metabolism [71], to the thousands of reactions in a complex genome-scale metabolism. FBA uses information about the stoichiometry of each reaction to predict steady state fluxes for all reactions in a metabolic network. The necessary stoichiometric information is represented as a stoichiometric matrix, S, of dimensions $m \times n$, where m denotes the number of metabolites, and n denotes the number of reactions in a metabolism [34,71]. FBA assumes that the concentrations of intracellular metabolites are in a steady state, which allows one to impose the constraint of mass conservation on them. This constraint can be written as $Sv = 0$, where v denotes a vector of metabolic fluxes through each reaction in a metabolism. The above equation has a large space of possible solutions, but not all of these solutions may be of biological interest. To restrict this space to fluxes of interest, FBA uses linear programming to maximize a biologically relevant quantity in the form of a linear objective function Z [71]. Specifically, the linear programming formulation of an FBA problem can be expressed as

$$\max \mathbf{Z} = \max \{c^{T}v \mid Sv = 0, a \leq v \leq b\}$$

The vector c contains the set of scalar coefficients that represent the maximization criterion. The individual entries of vectors a and b, respectively, contain the minimal and maximally possible fluxes for each reaction in v. Irreversible reactions can only have fluxes with positive signs, whereas irreversible reactions can have fluxes of both signs.

We are here interested in predicting whether a metabolism can sustain life in a given spectrum of environments, that is, whether it can synthesize all necessary small biomass molecules (biomass precursors) required for survival and growth. For our analysis of central carbon metabolism, there are 13 such essential precursors (Table S1b [35]). For our analysis of genome scale metabolisms, we use all 63 [30] biomass precursors of *E. coli*, because most of them would be required in any free-living

organism. They include 20 proteinaceous amino acids, DNA and RNA nucleotide precursors, lipids and cofactors. We use these biomass precursors to define the objective function and the vector *c*. We employed the package CLP (1.4, Coin-OR; https://projects.coin-or.org/Clp) to solve linear programming problems.

Growth environments.

Along with the biomass composition and stoichiometric information about a metabolic network, computational predictions of viability require information about the chemical environments that contain the nutrients needed to synthesize biomass precursors. In our analysis of central carbon metabolism, we consider a minimal aerobic growth environment composed of a sole carbon source, along with ammonium as a nitrogen source, inorganic phosphate as a source of phosphorus, as well as oxygen, protons, and water. When studying the viability of metabolisms on different carbon sources, we vary the carbon source while keeping all the other nutrients constant. When we say a particular metabolism is viable on 10 carbon sources, we mean that it can synthesize all biomass precursors when each of these carbon sources is provided as the sole carbon source in a minimal medium. The ten carbon sources we consider are D-glucose, acetate, pyruvate, D-lactate, D-fructose, alpha-ketoglutarate, fumarate, malate, succinate and glutamate.

Our analysis of genome-scale metabolisms requires a minimal environment with more nutrients, i.e., a sole carbon source, ammonium, inorganic phosphate, sulphate, sodium, potassium, cobalt, iron (Fe2+ and Fe3+), protons, water, molybdate, copper, calcium, chloride, magnesium, manganese and zinc [30]. For our analysis of genome-scale metabolisms viable on 10 carbon sources, we used the 10 carbon sources from the preceding paragraph.

The reactions used in the analysis of central carbon metabolism.

We use a global set of reactions in central carbon metabolism, which is based on a published reconstruction of *E. coli* central carbon metabolism [35]. From the published reconstruction [35], we deleted four reactions involved in ethanol synthesis, metabolism

and transport. We also grouped the reactions catalyzed by aconitase A and aconitase B into one reaction. We did this mainly to reduce the size of the set of reactions, in order to render the exploration of all variant metabolisms derived from it feasible. The final reaction set consists of $N = 51$ intracellular reactions, and we analyzed the viability of metabolisms comprising all possible 2^{51} subsets of this set. The reconstruction in [35] also involves 20 transport reactions, which are necessary to import nutrients or excrete waste products, and which we assume to be present in all metabolisms we studied.

The known reaction "universe" and the global metabolism.

We refer to the known universe of biochemical reactions as the set of reactions known to occur in some organism based on currently available biochemical knowledge. To arrive at this set, we curated data from the LIGAND database [31,32] of the Kyoto Encyclopedia of Genes and Genomes [72], which is divided into two smaller databases, the REACTION database and the COMPOUND database. These two databases together provide information about metabolic reactions, participating chemical compounds, and associated stoichiometric information. As described previously [10,11,13,33], we curated reactions from these databases by excluding reactions involving polymer metabolites of unspecified numbers of monomers; general polymerization reactions with uncertain stoichiometry; reactions involving glycans, owing to their complex structure; reactions with unbalanced stoichiometry; and reactions involving complex metabolites without detailed structural information [72]. After curation of these reactions, we added to them all non-redundant reactions from the published *E. coli* metabolic model (*i*AF1260), which comprises 1,397 non-transport reactions [30]. At the end of this procedure, we had arrived at a set of 5,906 non-transport reactions and 5,030 metabolites. We converted this set into what we call a global metabolism by including all *E. coli* transport reactions in this set [30]. Unsurprisingly, the global metabolism can synthesize all biomass precursors of *E. coli* from any of the carbon sources we consider here. We note that this metabolism may not be biologically realizable, for example, because it may contain thermodynamically infeasible pathways. However, we merely use it as a starting point to create smaller and ultimately minimal metabolisms through the sequential reaction deletion process described below.

Genotypes, phenotypes and viability.

The genes encoding the enzymes that catalyze a metabolism's reactions constitute the metabolic genotype of an organism. For our purpose, a more compact representation of a metabolic genotype is useful, which represents this genotype as a binary vector whose *i*-th entry corresponds to the *i*-th reaction in some global set or universe of biochemical reactions. This entry will be equal to one if an organism's genome encodes an enzyme capable of catalyzing this reaction, and zero otherwise (supplementary figure S1). The genotype space of all possible metabolisms comprises 2^N metabolisms, where N is the total number of known or considered chemical reactions ($N = 51$ for our analysis of central carbon metabolism, and $N = 5906$ for our analysis of genome-scale metabolisms). Any one organism's metabolic genotype can be thought of as a point in this space. Genotypes (metabolisms) viable in a given chemical environment are those that can synthesize all biomass precursors from nutrients in this environment.

Essential and nonessential reactions.

We define a reaction as essential for viability if its elimination abolishes viability in a given chemical environment. To identify all such essential reactions in a given metabolism, we eliminated each reaction and used FBA to assess whether non-zero biomass growth flux was still achievable. For our analysis of viability on 10 different (sole) carbon sources, we defined a reaction as essential if its elimination abolishes viability on *at least one* of the 10 carbon sources.

Identification of viable central carbon metabolisms.

To identify all viable metabolisms by exhaustive enumeration of viability for all 2^{51} (10^{15}) possible metabolisms in central carbon metabolism would be infeasible. Fortunately, such brute-force enumeration is also not necessary, for two reasons. The first originates from the notion of "environment-general superessential reactions" [13]. These are reactions whose elimination abolishes viability in each of the 10 carbon sources used here. To find such reactions, we converted the universe of central carbon metabolism into a format amenable to FBA analysis, as described earlier in this

section. We then deleted each reaction and determined viability on each of the 10 carbon sources. We found six reactions (supplementary table S1a, in red) that were necessary for biomass synthesis on each source. Any viable central carbon metabolism would require all six reactions, which reduces the number of metabolisms whose viability needs to be evaluated from 2^{51} to $2^{(51-6)} = 2^{45} (\approx 10^{13})$.

The second reason derives from a simple observation that reduces the number of genotypes whose viability needs to be determined even more dramatically: Removal of a reaction from an unviable metabolism cannot result in a viable metabolism. This means that among all metabolisms with *n*-1 reactions, we need to evaluate only the viability of those that are derived from viable metabolisms with *n* reactions through removal of one reaction. We incorporated this idea into an algorithm that allowed us to enumerate all viable genotypes [43].

Sampling of viable genome-scale metabolisms

To sample large (genome-scale) metabolisms, we started from the global metabolism of 5,906 reactions and deleted (eliminated) from it a sequence of randomly chosen reactions, while requiring that each such deletion preserves viability. Specifically, we chose a metabolic reaction at random and equiprobably among all reactions, deleted it, and used FBA to determine viability of the resulting metabolism. If the metabolism was viable, we accepted the deletion. Otherwise we randomly choose another reaction for deletion, and so on, until we found one whose deletion left the resulting metabolism viable. In this way, we deleted reactions until we had arrived at a minimal metabolism where no further reactions could be deleted. To determine whether a metabolism was minimal, we first attempted 1000 deletions of randomly chosen reactions, and if none of these deletions resulted in a viable metabolism, we deleted each reaction in the metabolism. If every one of these deletions resulted in nonviability, we declared the metabolism to be minimal.

Identification of viability-preserving paths connecting viable genotypes G_1 and G_2 at arbitrary size n.

To find out whether two genotypes G_1 and G_2 can be connected to one another through viability-preserving reaction swaps, we used the following heuristic approach. It does not rely on reaction swaps of arbitrary reactions, which we found to be too inefficient, but takes advantage of existing reactions in the two genotypes to accelerate the process. It defines a "walker" genotype G_1 and alters it through multiple random steps (reaction swaps) that approach the other, "target" genotype G_2. Before starting this random walk, we established two lists of reactions L_1 and L_2. L_1 contained all reactions in G_1 that were not contained in G_2. In this list we placed reactions non-essential in G_1 first (and in random order), followed by reactions essential in G_1 (also in random order). Conversely, L_2 consisted of arbitrarily ordered essential reactions in G_2, followed by arbitrarily ordered reactions nonessential in G_2. Each step in the random walk consisted of two parts, i.e., (i) adding to G_1 a reaction from L_2 (i.e., a reaction essential in G_2), and (ii) deleting from G_1 a reaction listed in L_1. Subsequent steps used subsequent reactions in each list for addition and deletion.

As this walk through genotype space progressed, we continued adding reactions to G_1 until all essential reactions from G_2 in list L_2 had been added to G_1, and continued from there on to adding nonessential reactions from L_2. During part (ii) of any given step, if none of the remaining reactions in the list could be deleted from walker G_1 without losing viability, we reverted the last reaction addition, and chose instead a reaction at random from the universe as a candidate for addition. Before adding it, we ensured that the chosen reaction shared all of its substrates and products with other reactions in the random walker. If the product of the reaction was not shared with another reaction, we checked if it could be secreted by a transport reaction. A candidate reactions that did not fulfill both criteria would be disconnected from the rest of metabolism, could therefore not possibly contribute to viability [33], and we discarded it, choosing another candidate, and so on, until we had found one that fulfilled both criteria. We then determined if, after the addition of this reaction, some reaction in the list could be deleted from the random walker. If so, we accepted the resulting swap, otherwise we tried another addition, and so on, until we had found an acceptable swap.

The two parts of each step ensure, first, that essential reactions from the target are preferentially added to the walker, thus increasing the likelihood of adding "useful" reactions to G_1, perhaps from one of several alternative metabolic pathways. Second, they reduce the chances of yielding an unviable genotype after a reaction deletion. However, the probability that the deletion of a reaction from walker G_1 can render it unviable increases with the number of reaction swaps, because past steps may have rendered previously nonessential reactions essential. We therefore also needed to use FBA after each deletion to ensure that G_1 retained viability after a reaction deletion.

We continued this guided random walk for as many swaps as needed to reach the target G_2, or until we had performed 5000 attempted swaps. In the latter case, we declared G_1 and G_2 disconnected. We note that this is no proof of disconnectedness, as some path may exist that this procedure cannot find. However, in practice, all our attempts to connect genotype pairs in this way were successful.

Identification of a metabolism's viable neighbors.

Two metabolisms are adjacent or neighbors of each other with respect to a reaction swap if they differ by one such swap. If the focal metabolism contains n reactions, then there are N-n reactions that are not part of the focal metabolism, where N is the total number of reactions a metabolism could possibly have. One can thus obtain a neighbor of the focal metabolism by deleting one of its n reactions and simultaneously adding one of the N-n reaction from the universe of reactions. Any metabolism therefore has $n \times (N\text{-}n)$ possible neighbors. To identify the viable neighbors of a minimal metabolism, we generated all possible $n \times (N\text{-}n)$ neighbors, and used FBA to determine their viability on glucose. (We also note that any minimal metabolism trivially has zero viable neighbors with respect to reaction deletion, and N-n viable neighbors with respect to reaction additions).

We used MATLAB (Mathworks Inc.) for all numerical analysis. Genotype space visualization was generated using the script available at (http://www.oaslab.com/Drawing_funnels.html).

4.6 Supplementary results

The connectivity of central carbon metabolisms with $n \geq 29$ reactions

We used the software `igraph`[45] to determine connectedness of genotype networks for metabolisms of size 23-28 and 46-50. We found that these genotype networks are mostly connected, as discussed in the main text. Unfortunately, that approach proved computationally too demanding for exploring the connectedness of metabolisms of greater sizes. We thus used another method to assess whether genotype networks of metabolisms at intermediate sizes are connected. At the heart of our method is the observation that if some subgraph A of a genotype network containing genotypes of size n is connected, then the genotypes $G(n+1)$ obtained from A by a single reaction addition also form a connected set. To illustrate our approach for small values of n, we start with the two components in the genotype network of size $n = 28$. Supplementary figure S3A displays the two components in the genotype network corresponding to size 28, one containing 434234 genotypes, and the other containing just 4 genotypes. By adding one reaction to each of these genotypes, we generate metabolisms of size 29. The two corresponding sets have 1773853 and 88 genotypes respectively (supplementary figure S3B), in addition to the remainder of 28 viable metabolisms of size 29. These 28 metabolisms are minimal in nature. Because these genotypes are minimal, they cannot belong to the two sets that were generated by the reaction addition process. Taking these new genotypes into account, one arrives at the total number of viable genotypes in $V(29)$ (supplementary figure S3B and table 4.1).

The arguments in the main text (section "Determining connectedness of genotype networks") show why each set obtained by reaction addition must be connected, but they do not tell us whether these two components are connected to one another, nor do they not inform us how the 28 minimal metabolisms connect to them. To determine their connectedness, we first checked if the 28 minimal metabolisms form a connected set using `igraph`[45]. We repeated the procedure and found that these 28 minimal metabolisms and the set of 88 metabolisms are also connected, yielding a larger component of size 116 (supplementary figure S3B). We next checked if any one of these 116 genotypes is separated by a reaction swap (constituting a neighbor) from another genotype in a set of 10000 sampled genotypes from the larger

component of 1773853 genotypes. We found many pairs that were neighbors, demonstrating that metabolisms of size 29 form one fully connected genotype network.

The analysis can be extended to larger sizes. By our argument in the main text, all genotypes of size 30 that derive from genotypes of size 29 by addition of one reaction form a connected set. In addition to these 5900563 genotypes (supplementary figure S3C), metabolisms at size 30 also contain 15 new minimal metabolisms (supplementary figure S3C and table 4.1). We examined them in the same way as described above, and found them all to be connected to the much larger component. In other words, all viable metabolisms with $n = 30$ reactions form a single connected component. At each step thereafter in this recursive reasoning to go from n to n+1, it is enough to determine all minimal metabolisms at size n+1 and to verify that they are neighbors of genotypes obtained by one reaction addition to viable genotypes at size n.

Essential pathways cause genotype network fragmentation

We here discuss another example that illustrates why pathway essentiality can lead to genotype network fragmentation. This one involves genotypes from the large connected component of size 25 (subgraphs *A''* and *B''* in figure 4.3C), on the one hand, and component *C* (blue in figure 4.3C), on the other. Again, we first calculated the superessentiality indices of all reactions for these components, and examined which reactions differ in their superessentiality index between the two components. We find that reactions catalyzed by phosphofructokinase (PFK), fructose-biphosphate aldolase (FBAl), and triose phosphate isomerase (TPI) are essential in all genotypes in the large component (supplementary figure S4, in green). Conversely, reactions catalyzed by phophoenolpyruvate synthase (PPS) and adenylate kinase (ADK1) (blue, supplementary figure S4) are essential in all four genotypes belonging to component *C* (figure 4.3C). For the purpose of illustration, supplementary figure S4 indicates in the form of one figure, the differences between a pair of genotypes (G_1, G_2), where G_1 belongs to the large component in figure 4.3C (subgraphs *A''* and *B''*, genotype includes reactions in green, but does not contain reactions in blue – see supplementary

figure S4) and the other genotype G_2 belongs to component *C* in figure 4.3C (genotype includes reactions in blue, but does not contain reactions in green – see supplementary figure S4). Reactions in black are encoded by both genotypes G_1 and G_2. We chose these specific genotypes G_1 and G_2 because they are separated by three reaction swaps, which is also the minimal distance between a pair of metabolisms of size 25 belonging to different components. Unlike the example discussed in the main text, where the essential pathways formed alternative routes (figure 4.4), these two essential pathways do not form alternative routes towards synthesizing the same molecules.

In addition to reactions catalyzed by PPS and ADK1, genotype G_1 does not contain the reaction catalyzed by TKT1, and is thus unable to synthesize sedeheptulose-7-phosphate (s7p). Though the reaction catalyzed by TALA is present in G_1 (as well as in G_2), it is nonfunctional in G_1 as s7p is one of its substrates. This also means that ribose-5-phosphate (r5p) and erythrose-4-phosphate (e4p) are synthesized through the pathway catalyzed by reactions G6PDH, PGL and GND (supplementary figure S4) as described in the example in the main text.

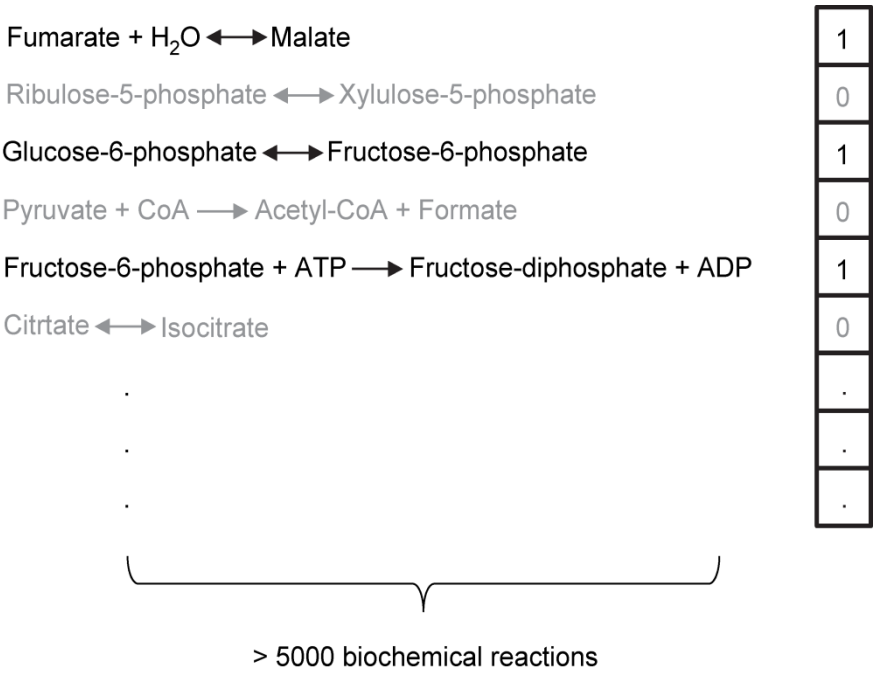
In genotype G_2, the absence of reactions PFK, FBAl and TPI (supplementary figure S4) forces all carbon that enters central carbon metabolism in the form of glucose-6-phosphate (g6p) towards the pentose phosphate pathway via the reaction catalyzed by glucose-6-phosphate dehydrogenase (G6PDH). Flux balance analysis shows that this results in most of the carbon being secreted as carbon dioxide (reaction GND (supplementary figure S4)), while a minority of it can enter the lower part of glycolysis as g3p and is converted further through GAPD, PGK, PGM and ENO into phosphoenolpyruvate (pep, supplementary figure S4). Secretion of a majority of carbon in the form of carbon dioxide results in the flux in the lower half of glycolysis to be smaller than the flux at which glucose-6-phosphate is generated in genotype G_2. However, pep needs to be generated at the same rate as glucose is imported through the glucose phosphotransferase system (GLCPTS) to maintain steady state. Because the flux in the lower half of glycolysis is smaller than the upper half, there is a flux deficit in the rate at which phosphoenolpyruvate is synthesized.

This flux deficit is rectified by phosphoenolpyruvate synthase (PPS). PPS is thus necessary in G_2 for the biosynthesis of the required amount of phosphoenolpyruvate

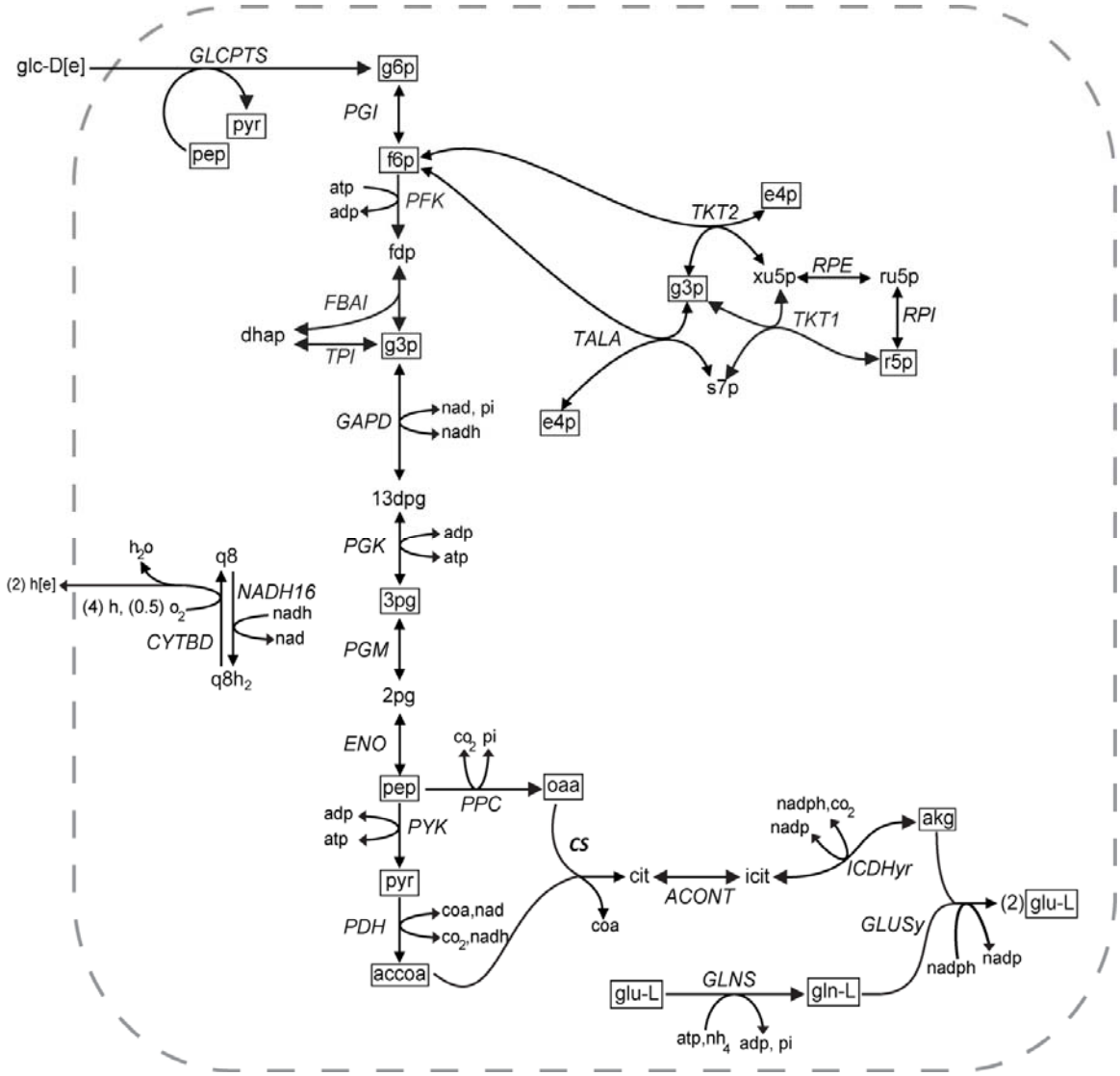
(pep). While PPS fulfills these demands, it also generates adenosine monophosphate (AMP) as a by-product. To ensure mass balance, this AMP needs to be utilized, which can be accomplished by adenylate kinase (ADK1) (supplementary figure S4). The reaction that ADK1 catalyzes is, in fact, the only reaction in our central carbon metabolism model that utilizes AMP. Thus, the requirement of phosphoenolpyruvate renders reactions PPS and ADK1 essential. Furthermore, because reactions catalyzed by PFK, FBAl and TPI are absent in G_2 reactions catalyzed by G6PDH, PGL, GND, TKT1 and TALA are all essential in this genotype.

In light of these observations, one also might ask why genotype networks of metabolisms with 25 and 26 reactions are disconnected (figure 4.2A and figure 4.3C), while genotype networks of parent metabolisms with 27 reactions are connected (figure 4.2A). As mentioned above, genotypes G_1 and G_2 are of size 25 and separated by three reaction swaps. Reactions catalyzed by PFK, FBAl and TPI are present in genotype G_1 and not in G_2, while reactions PPS, ADK1 and TKT1 are present in genotype G_2, but not in G_1. To generate parent metabolisms of size 26 from G_1 and G_2, we could add PPS to G_1 resulting in G_1', while adding PFK to G_2 would result in G_2'. Genotypes G_1' and G_2' would now be separated by two reaction swaps. Repeating this process by addition of ADK1 to G_1' and FBAl to G_2' would result in genotypes G_1'' and G_2'' of size 27, which are separated by one reaction swap and thus connected. That is, G_1'' could now be converted to G_2'' by addition of TKT1 and removal of TPI (supplementary figure S4).

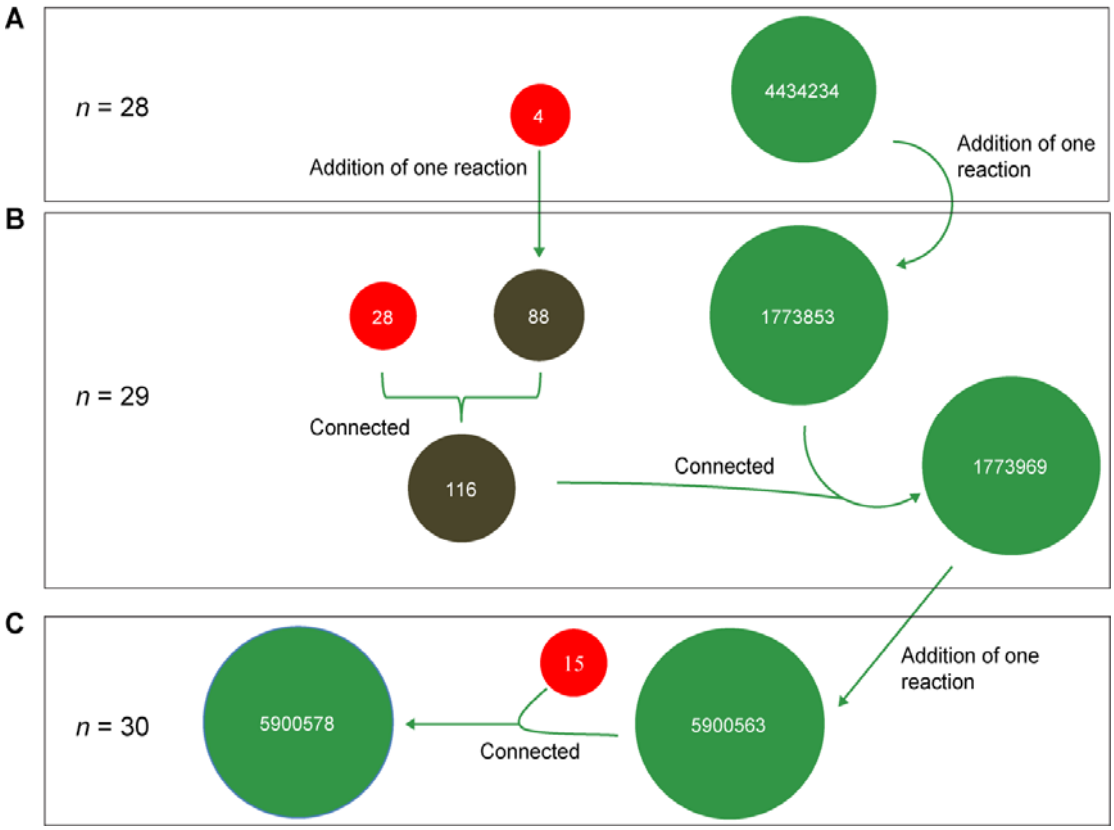
Supplementary figures



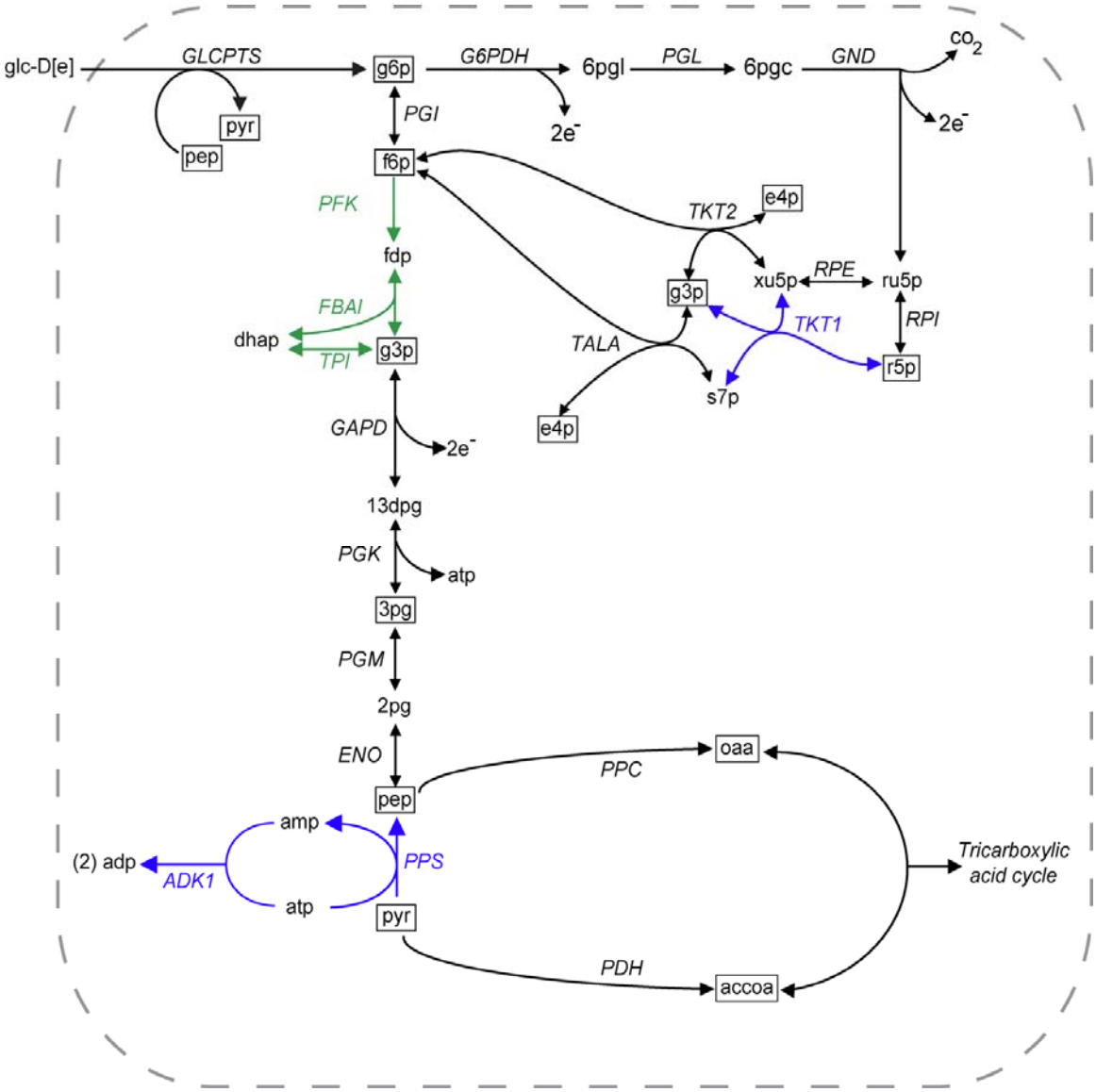
Supplementary figure S1: Representation of a genotype vector. Any genotype encoding n reactions ($n \leq N$) can be represented as a binary vector of length N, with n entries equal to one and all others equal to zero. The reactions that are present in the above hypothetical genotype are shown in black and the reactions that are absent are shown in grey.



Supplementary figure S2: An example of a minimal metabolism viable on glucose. The figure shows an example of a minimal metabolism of size 23, which is also one of the smallest metabolisms viable on glucose. All 13 biomass precursors are framed with solid rectangles. Only important transport reactions and cofactors are shown. Enzymes catalyzing each of the reactions are shown in uppercase italic typeface. Abbreviations are spelled out in supplementary table S1.



Supplementary figure S3: Connectedness of genotype networks containing viable genotypes of n = 28, 29, and 30 reactions. The figure shows the connectivity of genotype networks for reactions in central carbon metabolisms, as a function of size n. Each circle corresponds to a connected component, and the number in each circle corresponds to the number of genotypes in this component. The components at size 28 were obtained by full enumeration, but for larger sizes such an approach is not feasible. Instead one has to use a form of recursive evaluation that we illustrate here for two larger sizes. Panel (A) shows the two disconnected components in the network corresponding to size 28, one containing 434234 genotypes, and the other component containing just 4 genotypes. The addition of one reaction to these 4 genotypes results in 88 genotypes of size 29, which must be connected (see main text). (B) Aside from these 88 connected genotypes, there are also 28 minimal genotypes of size 29 (red circles). We verified computationally that both groups of genotypes (88 and 29) were connected using breadth-first search and found that they form a single component of 116 genotypes. We were also able to demonstrate that this component is connected to the 1773853 connected genotypes that are parents of the large component at size n = 28 (panel A). The two thus form a connected genotype network of 1773969 metabolisms of size 29. Panel (C) shows that adding one reaction to these genotypes results in 5900563 connected genotypes at size 30. In addition, 15 new minimal metabolisms (red circles) come into being at size 30. We found that they were connected to the remaining 5900563 genotypes, thus forming a single connected network comprising 5900578 genotypes.



Supplementary figure S4: An example of essential pathways that result in fragmentation of genotype space. The figure shows the two essential pathways (green and blue) in a pair of genotypes belonging to the two disconnected components in the genotype network of metabolisms with $n = 25$ reactions (figure 4.3C). The reactions in green, catalyzed by enzymes PFK, FBAl and TPI are essential in all genotypes belonging to the largest component (subgraphs *A''* and *B''*) in figure 4.3C, while the reactions in blue are essential to all four genotypes in component *C* in figure 4.3C. The 13 biomass precursors are surrounded by black rectangles. Only important transport reactions and cofactors have been shown. Enzymes catalyzing each of the reactions are shown in uppercase italic typeface. Information on abbreviations is provided in supplementary table S1.

Supplementary tables

Supplementary table S1a. This table lists the reactions of central carbon metabolism considered in this study. The suffix [e] denotes extracellular location of the metabolite. Rows in red correspond to environment-general superessential reactions.

Reaction abbreviation	Enzyme name	Reaction equation
ICL	Isocitrate lyase	icit --> glx + succ
MALS	malate synthase	accoa + glx + h2o --> coa + h + mal-L
ME1	malic enzyme (NAD)	mal-L + nad --> co2 + nadh + pyr
ME2	malic enzyme (NADP)	mal-L + nadp --> co2 + nadph + pyr
PPC	phosphoenolpyruvate carboxylase	co2 + h2o + pep --> h + oaa + pi
PPCK	phosphoenolpyruvate carboxykinase	atp + oaa --> adp + co2 + pep
ACONT	aconitase	cit <==> icit
AKGDH	2-Oxogluterate dehydrogenase	akg + coa + nad --> co2 + nadh + succoa
CS	citrate synthase	accoa + h2o + oaa --> cit + coa + h
FUM	fumarase	fum + h2o <==> mal-L
ICDHyr	isocitrate dehydrogenase (NADP)	icit + nadp <==> akg + co2 + nadph
MDH	malate dehydrogenase	mal-L + nad <==> h + nadh + oaa
SUCOAS	succinyl-CoA synthetase (ADP-forming)	atp + coa + succ <==> adp + pi + succoa
GLNS	glutamine synthetase	atp + glu-L + nh4 --> adp + gln-L + h + pi
GLUDy	glutamate dehydrogenase (NADP)	glu-L + h2o + nadp <==> akg + h + nadph + nh4
GLUN	glutaminase	gln-L + h2o --> glu-L + nh4
GLUSy	glutamate synthase	akg + gln-L + h + nadph --> (2) glu-L

	(NADPH)	+ nadp
ENO	enolase	2pg <==> h2o + pep
FBAl	fructose-bisphosphate aldolase	fdp <==> dhap + g3p
FBP	fructose-bisphosphatase	fdp + h2o --> f6p + pi
GAPD	glyceraldehyde-3-phosphate dehydrogenase	g3p + nad + pi <==> 13dpg + h + nadh
PDH	pyruvate dehydrogenase	coa + nad + pyr --> accoa + co2 + nadh
PFK	phosphofructokinase	atp + f6p --> adp + fdp + h
PGI	glucose-6-phosphate isomerase	g6p <==> f6p
PGK	phosphoglycerate kinase	3pg + atp <==> 13dpg + adp
PGL	6-phosphogluconolactonase	6pgl + h2o --> 6pgc + h
PGM	phosphoglycerate mutase	2pg <==> 3pg
PPS	phosphoenolpyruvate synthase	atp + h2o + pyr --> amp + (2) h + pep + pi
PYK	pyruvate kinase	adp + h + pep --> atp + pyr
TPI	triose-phosphate isomerase	dhap <==> g3p
ADK1	adenylate kinase	amp + atp <==> (2) adp
ATPM	ATP maintenance requirement	atp + h2o --> adp + h + pi
ATPS4r	ATP synthase (four protons for one ATP)	adp + (4) h[e] + pi <==> atp + (3) h + h2o
CYTBD	cytochrome oxidase bd (ubiquinol-8: 2 protons)	(2) h + (0.5) o2 + q8h2 --> (2) h[e] + h2o + q8
FRD7	fumarate reductase	fum + q8h2 --> q8 + succ
NADH16	NADH dehydrogenase (ubiquinone-8 & 3 protons)	(4) h + nadh + q8 --> (3) h[e] + nad + q8h2
NADTRHD	NAD transhydrogenase	nad + nadph --> nadh + nadp
SUCDi	succinate dehydrogenase (irreversible)	q8 + succ --> fum + q8h2

THD2	NAD(P) transhydrogenase	(2) h[e] + nadh + nadp --> (2) h + nad + nadph
G6PDH	glucose 6-phosphate dehydrogenase	g6p + nadp <==> 6pgl + h + nadph
GND	phosphogluconate dehydrogenase	6pgc + nadp --> co2 + nadph + ru5p-D
RPE	ribulose 5-phosphate 3-epimerase	ru5p-D <==> xu5p-D
RPI	ribose-5-phosphate isomerase	r5p <==> ru5p-D
TAL	transaldolase	g3p + s7p <==> e4p + f6p
TKT1	transketolase	r5p + xu5p-D <==> g3p + s7p
TKT2	transketolase	e4p + xu5p-D <==> f6p + g3p
ACALD	acetaldehyde dehydrogenase (acetylating)	acald + coa + nad <==> accoa + h + nadh
ACKr	acetate kinase	ac + atp <==> actp + adp
LDH_D	D-lactate dehydrogenase	lac-D + nad <==> h + nadh + pyr
PFL	pyruvate formate lyase	coa + pyr --> accoa + for
PTAr	phosphotransacetylase	accoa + pi <==> actp + coa
ACALDt	acetaldehyde reversible transport	acald[e] <==> acald
ACt2r	acetate reversible transport via proton symport	ac[e] + h[e] <==> ac + h
AKGt2r	2-oxoglutarate reversible transport via symport	akg[e] + h[e] <==> akg + h
CO2t	CO2 transporter via diffusion	co2[e] <==> co2
D_LACt2	D-lactate transport via proton symport	h[e] + lac-D[e] <==> h + lac-D
FORt2	formate transport via proton symport (uptake only)	for[e] + h[e] --> for + h
FORti	formate transport via	for --> for[e]

	diffusion	
FRUpts2	Fructose transport via PEP:Pyr PTS (f6p generating)	fru[e] + pep --> f6p + pyr
FUMt2_2	Fumarate transport via proton symport (2 H)	fum[e] + (2) h[e] --> fum + (2) h
GLCpts	D-glucose transport via PEP:Pyr PTS	glc-D[e] + pep --> g6p + pyr
GLNabc	L-glutamine transport via ABC system	atp + gln-L[e] + h2o --> adp + gln-L + h + pi
GLUt2r	L-glutamate transport via proton symport, reversible (periplasm)	glu-L[e] + h[e] <==> glu-L + h
H2Ot	H2O transport via diffusion	h2o[e] <==> h2o
MALt2_2	Malate transport via proton symport (2 H)	(2) h[e] + mal-L[e] --> (2) h + mal-L
NH4t	ammonia reversible transport	nh4[e] <==> nh4
O2t	o2 transport via diffusion	o2[e] <==> o2
PIt2r	phosphate reversible transport via proton symport	h[e] + pi[e] <==> h + pi
PYRt2r	pyruvate reversible transport via proton symport	h[e] + pyr[e] <==> h + pyr
SUCCt2_2	succinate transport via proton symport (2 H)	(2) h[e] + succ[e] --> (2) h + succ
SUCCt3	succinate transport out via proton antiport	h[e] + succ --> h + succ[e]

Biomass	Biomass Objective Function with GAM (growth-associate maintainence)	(1.496) 3pg + (3.7478) accoa + (59.8100) atp + (0.3610) e4p + (0.0709) f6p + (0.1290) g3p + (0.2050) g6p + (0.2557) gln-L + (4.9414) glu-L + (59.8100) h2o + (3.5470) nad + (13.0279) nadph + (1.7867) oaa + (0.5191) pep + (2.8328) pyr + (0.8977) r5p --> (59.8100) adp + (4.1182) akg + (3.7478) coa + (59.8100) h + (3.5470) nadh + (13.0279) nadp + (59.8100) pi

Supplementary table S1b. This table lists the metabolites of central carbon metabolism considered in this study. Rows in red correspond to biomass precursors in the central carbon metabolism. atp, nadph and nad are also biomass precursors, but we wish to emphasize on metabolites that are act as biochemical precursors to the actual biomass precursors of *E. coli* [30] (See main text).

Abbreviation	Metabolite full name
13dpg	3-Phospho-D-glyceroyl phosphate
2pg	D-Glycerate 2-phosphate
3pg	3-Phospho-D-glycerate
6pgc	6-Phospho-D-gluconate
6pgl	6-phospho-D-glucono-1,5-lactone
ac	Acetate
ac[e]	Acetate (extracellular)
acald	Acetaldehyde
acald[e]	Acetaldehyde (extracellular)
accoa	Acetyl-CoA
actp	Acetyl phosphate
adp	ADP
akg	2-Oxoglutarate
akg[e]	2-Oxoglutarate (extracellular)
amp	AMP
atp	ATP
cit	Citrate
co2	CO2
co2[e]	CO2 (extracellular)
coa	Coenzyme A
dhap	Dihydroxyacetone phosphate
e4p	D-Erythrose 4-phosphate
f6p	D-Fructose 6-phosphate
fdp	D-Fructose 1,6-bisphosphate
for	Formate
for[e]	Formate (extracellular)

fru[e]	D-Fructose (extracellular)
Fum	Fumarate
fum[e]	Fumarate (extracellular)
g3p	Glyceraldehyde 3-phosphate
g6p	D-Glucose 6-phosphate
glc-D[e]	D-Glucose (extracellular)
gln-L	L-Glutamine
gln-L[e]	L-Glutamine (extracellular)
glu-L	L-Glutamate
glu-L[e]	L-Glutamate (extracellular)
glu-L[e]	L-Glutamate (extracellular)
Glx	Glyoxylate
H	H+
h[e]	H+ (extracellular)
h2o	H2O
h2o[e]	H2O (extracellular)
Icit	Isocitrate
lac-D	D-Lactate
lac-D[e]	D-Lactate (extracellular)
mal-L	L-Malate
mal-L[e]	L-Malate (extracellular)
Nad	Nicotinamide adenine dinucleotide
Nadh	Nicotinamide adenine dinucleotide (reduced)
Nadp	Nicotinamide adenine dinucleotide phosphate
Nadph	Nicotinamide adenine dinucleotide phosphate (reduced)
nh4	Ammonium
nh4[e]	Ammonium (extracellular)
o2	O2
o2[e]	O2 (extracellular)
Oaa	Oxaloacetate
Pep	Phosphoenolpyruvate

pi	Phosphate
pi[e]	Phosphate (extracellular)
pyr	Pyruvate
pyr[e]	Pyruvate (extracellular)
q8	Ubiquinone-8
q8h2	Ubiquinol-8
r5p	alpha-D-Ribose 5-phosphate
ru5p-D	D-Ribulose 5-phosphate
s7p	Sedoheptulose 7-phosphate
succ	Succinate
succ[e]	Succinate (extracellular)
succoa	Succinyl-CoA
xu5p-D	D-Xylulose 5-phosphate

Supplementary table S2. Number of metabolisms viable on glucose with respect to metabolism size.

Metabolism size *n*	**Number of metabolisms Ω(*n*)**	**Number of viable metabolisms *V*(*n*) on glucose**	**Fraction of viable metabolisms on glucose**
23	1.9679E+14	3	1.52444E-14
24	2.2959E+14	91	3.96355E-13
25	2.4796E+14	1333	5.37588E-12
26	2.4796E+14	1.2512E+04	5.04599E-11
27	2.2959E+14	8.4344E+04	3.67365E-10
28	1.9679E+14	4.3424E+05	2.20657E-09
29	1.5608E+14	1.7740E+06	1.1366E-08
30	1.1446E+14	5.9006E+06	5.15529E-08
31	7.7535E+13	1.6274E+07	2.09889E-07
32	4.8459E+13	3.7712E+07	7.78212E-07
33	2.7901E+13	7.4145E+07	2.65744E-06

34	1.4771E+13	1.2456E+08	8.43246E-06
35	7.1745E+12	1.7967E+08	2.50429E-05
36	3.1887E+12	2.2325E+08	7.0013E-05
37	1.2927E+12	2.3933E+08	0.000185138
38	4.7626E+11	2.2138E+08	0.00046484
39	1.5875E+11	1.7647E+08	0.0011116
40	4.7626E+10	1.2087E+08	0.002537936
41	1.2778E+10	7.0817E+07	0.005542198
42	3.0423E+09	3.5259E+07	0.01158958
43	6.3676E+08	1.4786E+07	0.023220727
44	1.1578E+08	5.1600E+06	0.044569549
45	1.8009E+07	1.4745E+06	0.081871194
46	2.3491E+06	3.3744E+05	0.143647672
47	2.4990E+05	5.9966E+04	0.239959984
48	2.0825E+04	7909	0.379783914
49	1275	721	0.565490196
50	51	40	0.784313725
51	1	1	1

Supplementary table S3. Number of metabolisms viable on all ten carbon sources as a function of metabolism size.

Metabolism size *n*	**Number of metabolisms Ω(*n*)**	**Number of metabolisms viable on all carbon sources**	**Fraction of metabolisms viable on all carbon sources**
34	1.4771E+13	4	2.708E-13
35	7.1745E+12	79	1.10112E-11
36	3.1887E+12	736	2.30817E-10
37	1.2927E+12	4249	3.2869E-09
38	4.7626E+11	1.6869E+04	3.54197E-08
39	1.5875E+11	4.8507E+04	3.05549E-07

40	4.7626E+10	1.0399E+05	2.18349E-06
41	1.2778E+10	1.6902E+05	1.32277E-05
42	3.0423E+09	2.1017E+05	6.90827E-05
43	6.3676E+08	2.0060E+05	0.000315025
44	1.1578E+08	1.4667E+05	0.001266835
45	1.8009E+07	8.1539E+04	0.004527565
46	2.3491E+06	3.3985E+04	0.014467489
47	2.4990E+05	1.0377E+04	0.04152461
48	2.0825E+04	2237	0.107418968
49	1275	320	0.250980392
50	51	27	0.529411765
51	1	1	1

4.7 References

1. Schuster, P., Fontana, W., Stadler, P. F. & Hofacker, I. L. From sequences to shapes and back: a case study in RNA secondary structures. *Proc. Biol. Sci.* **255,** 279–84 (1994).

2. Reidys, C., Stadler, P. F. & Schuster, P. Generic properties of combinatory maps: neutral networks of RNA secondary structures. *Bull. Math. Biol.* **59,** 339–97 (1997).

3. Ancel, L. W. & Fontana, W. Plasticity, evolvability, and modularity in RNA. *J. Exp. Zool.* **288,** 242–83 (2000).

4. Wagner, A. Robustness and evolvability: a paradox resolved. *Proc. Biol. Sci.* **275,** 91–100 (2008).

5. Dill, K. A., Ozkan, S. B., Shell, M. S. & Weikl, T. R. The protein folding problem. *Annu. Rev. Biophys.* **37,** 289–316 (2008).

6. Honeycutt, J. D. & Thirumalai, D. The nature of folded states of globular proteins. *Biopolymers* **32,** 695–709 (1992).

7. Ciliberti, S., Martin, O. C. & Wagner, A. Robustness can evolve gradually in complex regulatory gene networks with varying topology. *PLoS Comput. Biol.* **3,** e15 (2007).

8. Payne, J. L. & Wagner, A. Constraint and contingency in multifunctional gene regulatory circuits. *PLoS Comput. Biol.* **9,** e1003071 (2013).

9. Samal, A. & Jain, S. The regulatory network of E. coli metabolism as a Boolean dynamical system exhibits both homeostasis and flexibility of response. *BMC Syst. Biol.* **2,** 21 (2008).

10. Samal, A., Matias Rodrigues, J. F., Jost, J., Martin, O. C. & Wagner, A. Genotype networks in metabolic reaction spaces. *BMC Syst. Biol.* **4,** 30 (2010).

11. Matias Rodrigues, J. F. & Wagner, A. Evolutionary plasticity and innovations in complex metabolic reaction networks. *PLoS Comput. Biol.* **5,** e1000613 (2009).

12. Matias Rodrigues, J. F. & Wagner, A. Genotype networks, innovation, and robustness in sulfur metabolism. *BMC Syst Biol* **5,** 39 (2011).

13. Barve, A., Rodrigues, J. F. M. & Wagner, A. Superessential reactions in metabolic networks. *Proc. Natl. Acad. Sci. U. S. A.* **109,** E1121–30 (2012).

14. Jörg, T., Martin, O. C. & Wagner, A. Neutral network sizes of biological RNA molecules can be computed and are not atypically small. *BMC Bioinformatics* **9,** 464 (2008).

15. Reidhaar-Olson, J. F. & Sauer, R. T. Functionally acceptable substitutions in two alpha-helical regions of lambda repressor. *Proteins* **7,** 306–16 (1990).

16. Baba, T. *et al.* Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol. Syst. Biol.* **2,** 2006.0008 (2006).

17. Segrè, D., Vitkup, D. & Church, G. M. Analysis of optimality in natural and perturbed metabolic networks. *Proc. Natl. Acad. Sci. U. S. A.* **99,** 15112–7 (2002).

18. Schaper, S., Johnston, I. G. & Louis, A. A. Epistasis can lead to fragmented neutral spaces and contingency in evolution. *Proc. Biol. Sci.* **279,** 1777–83 (2012).

19. Fong, S. S., Joyce, A. R. & Palsson, B. Ø. Parallel adaptive evolution cultures of Escherichia coli lead to convergent growth phenotypes with different gene expression states. *Genome Res.* **15,** 1365–72 (2005).

20. Ibarra, R. U., Edwards, J. S. & Palsson, B. O. Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. *Nature* **420,** 186–9 (2002).

21. Newman, M. *Networks : an introduction*. (Oxford University Press, 2010).

22. Wagner, A. *The origins of evolutionary innovations*. in press (Oxford University Press, USA, 2011).

23. Ciliberti, S., Martin, O. C. & Wagner, A. Innovation and robustness in complex regulatory gene networks. *Proc. Natl. Acad. Sci. U. S. A.* **104,** 13591–6 (2007).

24. Huynen, M. A. Exploring phenotype space through neutral evolution. *J. Mol. Evol.* **43,** 165–9 (1996).

25. Fontana, W. & Schuster, P. Shaping space: the possible and the attainable in RNA genotype-phenotype mapping. *J. Theor. Biol.* **194,** 491–515 (1998).

26. Bornberg-Bauer, E. How are model protein structures distributed in sequence space? *Biophys. J.* **73,** 2393–403 (1997).

27. Aguirre, J., Buldú, J. M., Stich, M. & Manrubia, S. C. Topological structure of the space of phenotypes: the case of RNA neutral networks. *PLoS One* **6,** e26324 (2011).

28. Boldhaus, G. & Klemm, K. Regulatory networks and connected components of the neutral space. *Eur. Phys. J. B* **77,** 233–237 (2010).

29. Neidhardt, F. *Escherichia coli and salmonella : cellular and molecular biology*. (1996).

30. Feist, A. M. *et al.* A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol. Syst. Biol.* **3,** 121 (2007).

31. Goto, S., Okuno, Y., Hattori, M., Nishioka, T. & Kanehisa, M. LIGAND: database of chemical compounds and reactions in biological pathways. *Nucleic Acids Res.* **30,** 402–4 (2002).

32. Goto, S., Nishioka, T. & Kanehisa, M. LIGAND: chemical database of enzyme reactions. *Nucleic Acids Res.* **28,** 380–2 (2000).

33. Barve, A. & Wagner, A. A latent capacity for evolutionary innovation through exaptation in metabolic systems. *Nature* **500,** 203–6 (2013).

34. Price, N. D., Reed, J. L. & Palsson, B. Ø. Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nat. Rev. Microbiol.* **2,** 886–97 (2004).

35. Orth, J. D., Fleming, R. M. T. & Palsson, B. Ø. Reconstruction and Use of Microbial Metabolic Networks: the Core Escherichia coli Metabolic Model as an Educational Guide. *EcoSal Plus* **1,** (2010).

36. Meléndez-Hevia, E., Waddell, T. G., Heinrich, R. & Montero, F. Theoretical approaches to the evolutionary optimization of glycolysis--chemical analysis. *Eur. J. Biochem.* **244,** 527–43 (1997).

37. ENTNER, N. & DOUDOROFF, M. Glucose and gluconic acid oxidation of Pseudomonas saccharophila. *J. Biol. Chem.* **196,** 853–62 (1952).

38. Bar-Even, A., Flamholz, A., Noor, E. & Milo, R. Rethinking glycolysis: on the biochemical logic of metabolic pathways. *Nat. Chem. Biol.* **8,** 509–17 (2012).

39. Flamholz, A., Noor, E., Bar-Even, A., Liebermeister, W. & Milo, R. Glycolytic strategy as a tradeoff between energy yield and protein cost. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 10039–44 (2013).

40. Romano, A. H. & Conway, T. Evolution of carbohydrate metabolic pathways. *Res. Microbiol.* **147,** 448–55

41. Huynen, M. A., Dandekar, T. & Bork, P. Variation and evolution of the citric-acid cycle: a genomic perspective. *Trends Microbiol.* **7,** 281–291 (1999).

42. Noor, E., Eden, E., Milo, R. & Alon, U. Central carbon metabolism as a minimal biochemical walk between precursors for biomass and energy. *Mol. Cell* **39,** 809–20 (2010).

43. Hosseini, S.-R. Exhaustive genotype-phenotype mapping in metabolic genotype space. (2013). doi:10.3929/ethz-a-009999124

44. Hopcroft, J. & Tarjan, R. Algorithm 447: efficient algorithms for graph manipulation. *Commun. ACM* **16,** 372–378 (1973).

45. Csardi, G. & Nepusz, T. The igraph software package for complex network research. *InterJournal Complex Syst.* **1695,** 1695 (2006).

46. Bastian, M., Heymann, S. & Jacomy, M. Gephi: an open source software for exploring and manipulating networks. *ICWSM* **2,** 361–362 (2009).

47. Pál, C. *et al.* Chance and necessity in the evolution of minimal metabolic networks. *Nature* **440,** 667–70 (2006).

48. Bar-Even, A., Noor, E., Lewis, N. E. & Milo, R. Design and analysis of synthetic carbon fixation pathways. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 8889–94 (2010).

49. Meléndez-Hevia, E., Waddell, T. G. & Cascante, M. The puzzle of the Krebs citric acid cycle: assembling the pieces of chemically feasible reactions, and opportunism in the design of metabolic pathways during evolution. *J. Mol. Evol.* **43,** 293–303 (1996).

50. Mittenthal, J. E., Clarke, B., Waddell, T. G. & Fawcett, G. A new method for assembling metabolic networks, with application to the Krebs citric acid cycle. *J. Theor. Biol.* **208,** 361–82 (2001).

51. Ebenhöh, O. & Heinrich, R. Stoichiometric design of metabolic networks: multifunctionality, clusters, optimization, weak and strong robustness. *Bull. Math. Biol.* **65,** 323–57 (2003).

52. Giovannoni, S. J. *et al.* Genome streamlining in a cosmopolitan oceanic bacterium. *Science* **309,** 1242–5 (2005).

53. Dufresne, A., Garczarek, L. & Partensky, F. Accelerated evolution associated with genome reduction in a free-living prokaryote. *Genome Biol.* **6,** R14 (2005).

54. Akman, L. *et al.* Genome sequence of the endocellular obligate symbiont of tsetse flies, Wigglesworthia glossinidia. *Nat. Genet.* **32,** 402–7 (2002).

55. Pérez-Brocal, V. *et al.* A small microbial genome: the end of a long symbiotic relationship? *Science* **314,** 312–3 (2006).

56. Fraser, C. M. *et al.* The minimal gene complement of Mycoplasma genitalium. *Science* **270,** 397–403 (1995).

57. Fleischmann, R. D. *et al.* Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* **269,** 496–512 (1995).

58. Moran, N. & Wernegreen, J. Lifestyle evolution in symbiotic bacteria: insights from genomics. *Trends Ecol. Evol.* **15,** 321–326 (2000).

59. McCutcheon, J. P. & Moran, N. A. Parallel genomic evolution and metabolic interdependence in an ancient symbiosis. *Proc. Natl. Acad. Sci. U. S. A.* **104,** 19392–7 (2007).

60. Kikuchi, Y. *et al.* Host-symbiont co-speciation and reductive genome evolution in gut symbiotic bacteria of acanthosomatid stinkbugs. *BMC Biol.* **7,** 2 (2009).

61. Thomas, G. H. *et al.* A fragile metabolic network adapted for cooperation in the symbiotic bacterium Buchnera aphidicola. *BMC Syst. Biol.* **3,** 24 (2009).

62. Pál, C., Papp, B. & Lercher, M. J. Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. *Nat. Genet.* **37,** 1372–5 (2005).

63. Omelchenko, M. V, Makarova, K. S., Wolf, Y. I., Rogozin, I. B. & Koonin, E. V. Evolution of mosaic operons by horizontal gene transfer and gene displacement in situ. *Genome Biol.* **4,** R55 (2003).

64. Lawrence, J. G. & Roth, J. R. Evolution of coenzyme B12 synthesis among enteric bacteria: evidence for loss and reacquisition of a multigene complex. *Genetics* **142,** 11–24 (1996).

65. Boucher, Y. & Doolittle, W. F. The role of lateral gene transfer in the evolution of isoprenoid biosynthesis pathways. *Mol. Microbiol.* **37,** 703–16 (2000).

66. Bilgin, T. & Wagner, A. Design constraints on a synthetic metabolism. *PLoS One* **7,** e39903 (2012).

67. Fong, S. S., Marciniak, J. Y. & Palsson, B. Ø. O. Description and Interpretation of Adaptive Evolution of Escherichia coli K-12 MG1655 by Using a Genome-Scale In Silico Metabolic Model. *J. Bacteriol.* **185,** 6400–6408 (2003).

68. Vieira-Silva, S. & Rocha, E. P. C. The systemic imprint of growth and its uses in ecological (meta)genomics. *PLoS Genet.* **6,** e1000808 (2010).

69. Nam, H. *et al.* Network context and selection in the evolution to enzyme specificity. *Science* **337,** 1101–4 (2012).

70. Kim, J., Kershner, J. P., Novikov, Y., Shoemaker, R. K. & Copley, S. D. Three serendipitous pathways in E. coli can bypass a block in pyridoxal-5'-phosphate synthesis. *Mol. Syst. Biol.* **6,** 436 (2010).

71. Kauffman, K. J., Prakash, P. & Edwards, J. S. Advances in flux balance analysis. *Curr. Opin. Biotechnol.* **14,** 491–6 (2003).

72. Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M. & Hirakawa, M. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res.* **38,** D355–60 (2010).

PUBLICATIONS

- Barve A., Wagner A.: Historical contingency does not strongly constrain the gradual evolution of metabolic properties in central carbon and genome-scale metabolisms. *BMC Systems Biology* 8:48 (2014). [doi: 10.1186/1752-0509-8-48]
- Wagner A., Andriasyan V., Barve A.: The organization of metabolic genotype space facilitates adaptive evolution in nitrogen metabolism. *Journal of Molecular Biochemistry* 3.1 (2014).
- Barve A., Wagner A.: A latent capacity for evolutionary innovation through exaptation in metabolic systems. *Nature* 500, 203–206 (2013). [doi: 10.1038/nature12301]
- Sabath N., Ferrada E., Barve A., Wagner A.: Growth temperature and genome size in bacteria are negatively correlated, suggesting genome streamlining during thermal adaptation. *Genome Biology and Evolution* 5(5), 966-977 (2013). [doi: 10.1093/gbe/evt050]
- Barve A., Rodrigues J., Wagner A.: Superessential reactions in metabolic networks. *Proceedings of the National Academy of Sciences* 109 (18), E1121-E1130 (2012). [doi: 10.1073/pnas.1113065109]
- Barve A., Gupta A., Solapure S., Kumar A., Ramachandran V., Seshadri K., Vali S., Datta S.: A kinetic platform for in silico modeling of the metabolic dynamics of *Escherichia coli*. *Advances and Applications in Bioinformatics and Chemistry* 3, 97-110 (2010). [doi : 10.2147/AABC.S14368]

PROFESSIONAL EXPERIENCE

Cellworks Research India LTD., Bangalore, India

Biomodeling Scientist **December 2004 - August 2009**

- Infectious Disease
- Diabetes type 2